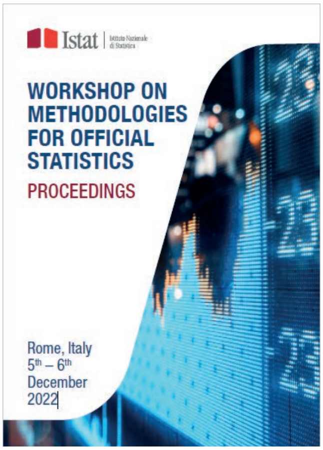
Istat Istituto Nazionale di Statistica

WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

PROCEEDINGS

Rome, Italy
5th – 6th
December
2022



Istat | Istituto Nazionale di Statistica

WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

PROCEEDINGS

Rome, Italy
5th – 6th
December
2022

Workshop on Methodologies for Official Statistics Proceedings

Programme Committee:
Orietta Luzi (coordinator)
Daniela Cocchi
Piero Demetrio Falorsi
Stefano Falorsi
Pierre Lavallée
Maurizio Lenzerini
Brunero Liseo
Maria Giovanna Ranalli
Monica Scannapieco
Mauro Scanu
Natalie Shlomo
Li-Chun Zhang

Editorial activities: Nadia Mignolli (coordinator), Alfredina Della Branca, Marco Farinacci, Alessandro Franzò e Manuela Marrone.

Responsible for graphics: Sofia Barletta.

ISBN 978-88-458-2102-8
© 2023
Istituto Nazionale di Statistica
Via Cesare Balbo, 16 – Roma

Index

Pag.

Welcome and opening session

SESSION 1

Methodologies for the new censuses

SESSION 2

Methodologies for multi-source processes

SESSION 3

Methodologies for big data

Pag.

SESSION 4

Standardisation of methods and processes

CLOSING SESSION

Welcome by the President of the Italian National Institute of Statistics - Istat

Gian Carlo Blangiardo

Welcome to all the online and present attendees in this first Istat Workshop on Methodologies for Official Statistics. This workshop has been organised in order to have a meeting place where peers can talk and debate on the latest issues on Statistical Methods applied in data production processes of National Statistical Institutes.

Istat is a research institute whose objective is the production of high-quality statistical information for policymakers and the whole society. This workshop, organised under the supervision and support of the Istat Advisory Committee on Statistical Methods, is one of the actions implemented to confirm our Institute’s commitment to quality, one of the key mandates stated in the European Statistics Code of Practice.

In this context, methodological research aims to define new objective methods applicable to the data production processes in order to improve the quality of statistical information. The fact that a statistical institute is an active contributor in the development of new statistical applications, not just a consumer of new methods emerging internationally, allows Istat to always aim at introducing the best approaches to improve the quality of statistical production. This strategy means that Istat continues to build trust and confidence on data production within the national and international scientific community, as well as among data users and stakeholders.

This workshop is a further demonstration of the importance given by Istat to the need to invest in the research of new methodological solutions to respond to ever-new and challenging needs in new and complex production contexts.

In this respect, it is important to underline that, even if the active presence of Istat researchers in the scientific community represents a long-standing tradition, internal methodological research has received a fundamental boost from some of the research infrastructures our Institute has been able to establish in more recent years.

The Advisory Committee on Statistical Methods represents one of these infrastructures.

Therefore, my sincere thankfulness and appreciation go to all the members of the Advisory Committee on Statistical Methods, who are leader statisticians working on different methodological research areas in academic institutions and National Statistical Institutes.

Special thanks go to the coordinator of the Committee, Professor Daniela Cocchi, retired from the University of Bologna in the last month, after a long and fruitful academic career.

I also wish to thank the other Committee members: Professor Natalie Shlomo (University of Manchester), Professor Maria Giovanna Ranalli (Università di Perugia), Mr Pierre Lavallée (formerly at Statistics Canada), Professor Li-Chun Zhang (University of Southampton and Statistics Norway), Professor Maurizio Lenzerini and Professor Brunero Liseo (Università La Sapienza di Roma), and Mr Piero Falorsi (formerly at Istat).

I take this opportunity to also thank all workshop discussants for accepting our invitation: Professor David Haziza (University of Ottawa), Mr Thomas Burg (Statistics Austria), Mr Piet Daas (Statistics Netherlands) and Mr Fabio Ricciato (Eurostat).

I wish everybody a very fruitful workshop.

The perspective of statistical production to the workshop

Monica Pratesi[1]

Istat is honored to host another Workshop on Methodologies for Official Statistics. This event has the objective to gather distinguished researchers on statistical Methodologies applied in the context of official statistics, to promote an exchange of ideas and good practices.

I am here to testify to the perspective of statistical production in the workshop. Every session of the workshop ends with a contribution from the point of view of the Statistical Production Department.

While research on Statistical Methods is certainly important and at the base of sound and innovative statistical production, ultimately it is the quality of the data produced and its relevance to real-world issues that will determine the success of statistical production processes. For these, the continuous interaction between the development of methods and the statistical production is crucial: I hope that the insights and experience offered to each session by researchers involved in the production will contribute to a productive and informative workshop.

My perspective is on the importance of practical experience and testing on the new methods proposed.

As you know, Statistical Methods are a relevant part of the overall statistical production process. High-quality data that is relevant to important issues can have a significant impact on policy decisions and other important outcomes. As such, it is important for statistical producers to keep the end users in mind when designing and implementing their processes, and to ensure that the data produced is of the highest quality possible. This may involve using a range of methods to ensure data accuracy, such as data cleaning, validation, and transformation, as well as thorough transparency and documentation of the data sources and methods used. Additionally, it is important to ensure that the data produced is easily accessible and understandable by end users, so that it can be used effectively to inform decision-making and drive positive outcomes.

[1] Monica Pratesi (monica.pratesi@istat.it), Italian National Institute of Statistics - Istat.

Yes, that is a correct interpretation of the phrase “the proof of the pudding is in the eating”. It means that the true value or quality of any process can only be determined by experiencing or using it, and not just by its appearance or reputation. The phrase is often used to emphasise the importance of practical experience and testing, rather than relying solely on theoretical or abstract knowledge.

During the statistical production process, often described by the well-known Generic Statistical Business Process Model (GSBPM), there are many production issues that can occur and concerns that arise: poor quality, timeliness (long lead times), high on-hand inventory («un-sold», that is not relevant or that exceeds the projected «users» demand), supply chain interruptions (interruption in the flow of the production process), etc.

All of these things affect the statistical products (data, indicators, analyses), which in turn affect the public’s perception of Official Statistics. The most common problems tend to fit into four categories:

- Quality problems: high defect rate (coverage, measurement, sampling error, etc.), low response rate, and poor quality.
- Output problems: long lead time, unreasonable production schedule, high inventory rate, supply chain interruption (interruption in the flow of process that involves any of the entities associated with the production).
- Cost problems: low efficiency, idle processes (persons – machines – technologies).
- Management problems: potential safety hazards (running risks, etc.), bad working conditions.

There are four sessions today and tomorrow, representing the Istat priority areas of research on Statistical Methods:

- methods for the new censuses;
- methods for multi-source processes;
- methods for big data;
- standardisation of methods and processes.

All of these methods are a possible response to the production issues related to maintaining quality, facilitating the production of outputs and controlling costs aimed at improving the management of the whole process.

In this workshop, the research activities and advancements in some of the most strategic areas of statistical production are addressed.

I want to underline the importance of two of them, as they represent the backbones of the new strategy for statistical production that has been progressively implemented in Istat thanks to the modernisation process launched in 2016.

- The area of multi-source processes and the Integrated System of Statistical Registers (ISSR): wide potentiality since different registers can be linked together on the basis of clearly defined keys.
- The area of Methodologies for the new Census and the Social Survey Integrated System: the final goal is the harmonisation between the three surveys (for Living Condition Survey, Labour Force Survey, EU-SILC Survey, and Household Budget Survey) and the Census Survey.

Integration between registers and surveys provides information to estimate target variables using suitable statistical models.

It is important to address these production issues with up-to-date statistical Methodologies and innovation as well as an adequate standardisation of methods and processes in order to maintain the integrity and accuracy of Official Statistics, as they can have a significant impact on public trust in data and the credibility of the institutions that produce them.

From the perspective of Istat's Department for Statistical Production, workshops like this one, are more than welcomed.

Thank you for your attention and my best wishes for a fruitful and pleasant experience in Istat.

Managing methodological research and innovation in Istat

Orietta Luzi[1]

The Italian National Institute of Statistics (Istat) is a public research body producing Official Statistics.

Methodological research is fundamental in Istat to ensure that the “best methods and practices” are used in statistical production processes, both traditional - such as direct sample surveys and censuses - and new ones - such as statistical registers based on integrated administrative archives and statistical processes based on big data and other non-traditional data sources.

To support the Istat strategy in the area of methodological research and innovation, some strategic assets have been created in recent years.

First of all a three-year strategic plan for research covering the period 2022-2024 has been approved (Istat, 2022). In the plan, the priority areas of investments for methodological and thematic research are indicated, so that human and financial resources could be primarily focussed on them.

Furthermore, Istat promotes and supervises research activities through some dedicated infrastructures, where experts from many different Directions and Units and also from external bodies collaborate. These infrastructures (Istat, 2023) are the Istat Research Committee, the Laboratories (the Innovation Laboratory and the Thematic Laboratory), and the Advisory Committee on Statistical Methods (MAC)[2], which has been active in Istat since 2017 and has been renewed in 2020.

The MAC members are National and International experts having both a refereeing and orientation role on specific research projects carried out by Istat researchers in the priority research areas, verifying their quality and alignment with the current state of research at international level. MAC members also propose and directly carry out advanced training courses for Istat researchers (the so-called master classes).

[1] Orietta Luzi (luzi@istat.it), Italian National Institute of Statistics - Istat.

[2] For further details see: https://www.istat.it/it/ricerca-in-istat/organizzazione/comitato-per-le-metodologie-statistiche.

It is important to underline the role of the MAC in very important in strengthening the cooperation among Istat and other National Statistical Institutes - NSIs and with (Italian and European) universities, to share experiences and new methodological solutions to common problems.

The Workshop on Methodologies for Official Statistics extensively exploits the last three-years' activity of the MAC, and at the same time is organised in four sessions corresponding to the Istat priority areas of methodological research for the 2022-2024 period:

1. Methodologies for the new permanent censuses;
2. Methodologies for multi-source processes;
3. Methodologies for big data;
4. Standardisation of Statistical Methods and processes.

In each session/research area, a first paper provides an overview of the research activities carried out and/or ongoing in Istat in this area, while a second paper deals with one of the projects discussed by the MAC in the same research area in the last three-year period.

Each session hosts two discussants: one expert "external" to Istat, coming from the University, from other NSIs, or Eurostat, and a second expert from the Istat Statistical Production Department, as it was considered very important to have the point of view of the internal statistical production areas on the research projects presented.

This workshop represents a first opportunity for sharing and discussing the research and methodological innovation directions pursued by Istat, and on future perspectives. The workshop will be a periodic appointment to continue this discussion and to stimulate collaboration between researchers of the Institute with researchers of other NSIs, as well as of the academic world.

The workshop organisation has been supported by a Programme Committee involving all the members of the MAC and all the members of the MAC secretariat, and by an Organisation Committee involving experts in the Communication and IT areas.

References

Istituto Nazionale di Statistica – Istat. 2022. *Piano triennale della ricerca tematica e metodologica*. Roma, Italy: Istat (Accessed on 7th March 2023). https://www.istat.it/it/files//2022/09/PianoTriennaleRicerca_15022022.pdf.

Istituto Nazionale di Statistica – Istat. 2023. "Organisation and research areas". Roma, *Organisation*. Italy: Istat (Accessed on 7th March 2023). https://www.istat.it/en/research-activity/organisation-.

SESSION 1

Methodologies for the new censuses

INTRODUCTION

Pierre Lavallée[1]

The Italian Permanent Census is a very important project for Istat. This new methodological framework has been designed and implemented starting from 2016. Its goal is to produce annual data — replacing the previous decennial cycle — using information from administrative sources integrated with sample surveys information. The Italian Permanent Census is register-based using: (i) the Integrated Register System (IRS); (ii) the Permanent Population Census; (iii) the Census and Social Survey Integrated System (CSSIS).

The target units of the Italian Permanent Census are the usual resident persons (living in a household). There are three classes of variables of interest: the register variables, the survey variables, and the non-replaceable variables. The register variables, mainly from administrative sources, include variables such as *sex, age, civil status*, and other variables like *educational level* and *occupational status*. The survey variables include variables such as *non-employment status and commuting*. These variables cannot be deduced from the administrative sources. The non-replaceable variables are not directly available from administrative data. For these variables, target parameters are estimated by means of sample surveys and exploiting the auxiliary information coming from the registers.

The IRS is the "Backbone" of the system for production of social statistics. The IRS production process is based on the of massive integration at single record level of multi-source of administrative and survey data. The IRS produces and manages the Population Register.

The Population Permanent Census provides fundamental information on the structure of the population, guaranteeing very high levels of territorial and sectoral granularity. It adds to the set of register variables the estimates from the sample surveys of variables that cannot be deduced from the administrative sources.

The CSSIS produces annual data for target parameters (so-called hypercubes), as well as multi-annual data for traditional parameters produced every 10 years. The CSSIS is used for filling information gap of the Population Register for estimation of target parameters.

[1] Pierre Lavallée (pierrelavallee_ca@yahoo.fr), formerly at Statistics Canada.

Due to its importance and complexity, several papers on the Italian Permanent Census have been presented to Istat's *Comitato Consultivo per le Metodologie Statistiche* (Advisory Committee on Statistical Methods). The Advisory Committee reviewed and commented these papers to solve issues, to improve and to enhance this interesting project. The following is a list of the various papers presented to the meetings of the Advisory Committee:

2017 (April)

- "Census and social survey integrated system", by Falorsi, S.
- "Balancing Methods for Ensuring Time and Space Consistency of Demographic Estimates in the Italian Integrated System of Statistical Registers", by Di Zio, M., M. Fortini, and D. Zardetto.
- "Integration of administrative sources and survey data through Hidden Markov Models for the production of labour statistics", by Guarnera, U., and D. Filipponi.

2017 (November)

- "The anticipated variance as a measure for the accuracy of complex multi-source statistics", by Righi, P., and P.D. Falorsi.

2019 (June)

- "Census and social survey integrated system (update)", by Falorsi, S.
- "A Hierarchical Bayesian model for quality check of the Italian population count by Administrative Data", by Toti, S., R.M. Lipsi, S. Giavante, and S. Daddi.

2019 (November)

- "The Italian Permanent Census and issues related to population counts estimation when data are affected by coverage error", by Fortini, M., S. Falorsi, and P. Righi.

- “A comparison between machine learning techniques and standard statistical methods for the imputation of the “attained level of education” in the base register of individuals”, by De Fausti, F., M. Di Zio, R. Filippini, S. Toti, and D. Zardetto.

2020 (June)

- “Census and social survey integrated system (update)”, by Falorsi, S., A. Bernardini, N. Cibella, M. D'Alò, L. Di Consiglio, M. Di Zio, A. Fasulo, D. Filipponi, M. Fortini, P. Righi, A. Ronconi, F. Solari, and S. Toti.
- “R package SamplingStrata: new developments and extension to Spatial Sampling”, by Ballin, M., and G. Barcaroli.
- “Imputation of the “Attained Level of Education” in Base Register of Individuals: a comparison between Machine Learning and standard techniques”, by De Fausti, F., R. Filippini, M. Di Zio, S. Toti, and D. Zardetto.
- “Current directions for research on record linkage in Istat: focus on Mixture models for probabilistic record linkage”, by Tuoto, T., and M. Fortini.

2021 (June)

- “Census count estimates geocoded at sub-domain levels”, by Daddi, S., M. Di Zio, M. D’Alò, S. Falorsi, and D. Filipponi.
- “A pseudo-population bootstrap approach for variance estimation of population counts with under/over coverage”, by Toti, S., M. Di Zio, and A. Ronconi.
- “LFS non-response indicators for register overcoverage estimation”, by Loriga, S., L. Di Consiglio, and S. Falorsi.

2021 (December)

- “Planning the Post-21 Permanent Census of Population and Housing according to a Responsive-Adaptive Survey Design approach”, by De Vitiis, C., S. Falorsi, A. Guandalini, F. Inglese, P. Righi, and M.D. Terribili.
- “A proposal for a spatial concentration index”, by Ballin, M.
- “Longitudinal and cross-sectional analyses of data in the Integrated System of Statistical Register”, by Altarocca, F., M.R. Aracri, R. Benedetti, R. Radini, and G. Vaste.

2022 (May)

- “Several data on labour status, a problem or a resource? Looking for an integrated approach for a good quality and consistent set of statistics”, by D’Alò, M., S. Falorsi, D. Filipponi, and S. Loriga.
- “A Statistical Framework for Register-Based Population Size Estimation”, by Bernardini, A., N. Cibella, and F. Solari.

It should be noted that many other papers written by Istat were not necessarily presented at the Advisory Committee. The Italian Permanent Census is a complex project where several issues and challenges are rising throughout its development.

At Session 1 of the first Workshop on Methodologies for Official Statistics, the following papers were presented:

- “Census and social survey integrated system” – presented by Stefano Falorsi.
- “Multi-source statistics in the Italian permanent census” – presented by Marco Di Zio.

The discussion was lead by Professor David Haziza (University of Ottawa, Canada) and the session terminated with the point of view of the Statistical Production Department of Istat, by Saverio Gazzelloni who is Head of Directorate for Social Statistics and Population Census.

Census and social survey integrated system

Michele D'Alò[1], *Stefano Falorsi*[1]

Abstract

Starting from October 2018, the population Census in Italy has abandoned the traditional decennial 'door-to-door' enumeration for a 'combined' approach which integrates administrative data and sample surveys. The goal of the 'permanent' Census is to produce annual data - replacing the previous decennial cycle - using information from administrative sources integrated with sample surveys information. The new Census strategy is planned to allow a significant reduction of the cost of the census, of respondents' burden, and of the organisational impact on municipalities.

Keywords: Master sample design, multi-source estimation, coverage errors, Multivariate small area estimators.

1. Introduction

Starting from October 2018, the population Census in Italy has abandoned the traditional decennial 'door-to-door' enumeration in favour of a 'combined' approach which integrates administrative data and sample surveys. In fact, in 2012, the so-called 'permanent' Census of Population and Housing (in Italian "Censimento permanente della popolazione e delle abitazioni") was introduced in Italian legislation (Article 3 of Legislative Decree 179/2012, converted with amendments into Law 221/2012). The 'permanent' Census besides ensuring the usual estimation of the ten years hypercubes on socio-economic variables required from Eurostat and for the production of others tables required to fulfil the national tabulation plan, allows also the computation of annual basic statistics at municipality level. This is done within the frame of Istat's (the Italian National Institute of Statistics) strategic programmes, whose focus is to integrate administrative data, create statistical registers and conduct supporting statistical surveys, in line with the new organisational, technological and methodological

[1] Michele D'Alò (dalo@istat.it); Stefano Falorsi (stfalors@istat.it), Italian National Institute of Statistics - Istat.

data production model aimed at fully exploiting all type of available data. The new Census strategy allows a significant reduction of the census' cost, of the respondents' burden and of the organisational aspects of municipalities' field - work, taking advantage of the results of several projects that since 2015, Istat lunched in order to exploit for statistical purposes the available administrative sources of information. A similar census design was studied by the UK Office for National Statistics, ONS (Office for National Statistics, 2016) and for the Israeli rolling integrated census (Pfeffermann, 2015).

2. The Italian Permanent Census

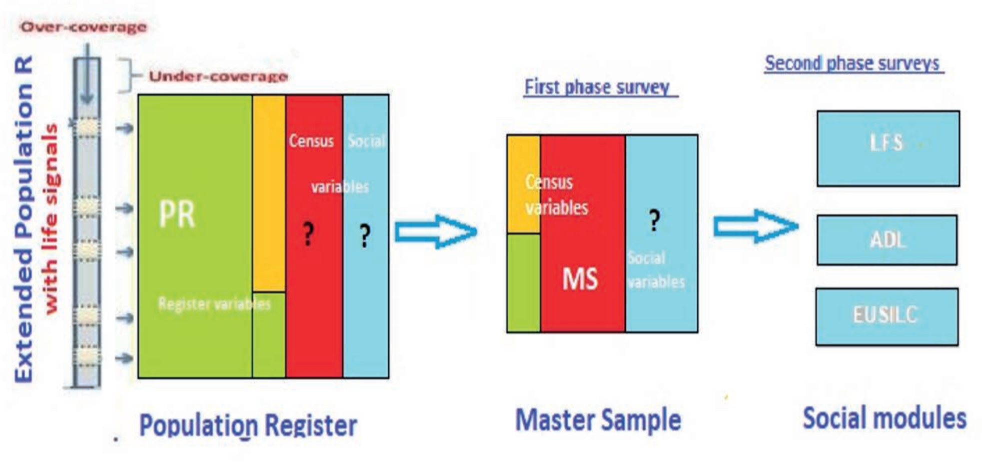
The strategy underlying the new Population Census System (PCS) aims to integrate the information stored into registers with those specifically collected through a specific "census" master sample survey. The Integrated System of Registers (ISR) is the backbone of the framework. It is built at single record level, mainly through a massive integration of administrative data, but also using information from surveys every time the subpopulations of interest are not covered by administrative data. Record linkage techniques, editing and imputation of missing data are then involved in this data production process. The set of variables derived from administrative sources are called register variables. They correspond to all the information, such as gender, age, marital status, that are primarily derived from demographic sources. For other information, such as the education level, the administrative information provides instead a good approximation for the large part of population records, except for some specific sub-populations. Finally, for information like employment status (employed/non-employed) administrative data are just strongly correlated with the target variables and for that, they are involved in a micro prediction stage aimed to compute the needed information for each record (Filipponi, Guarnera, Varriale, 2019 and Boeschoten, Filipponi, Varriale, 2021). The number of person belonging to the usual resident population has been initially computed using the information collected by the Census Master Sample (MS). Anyway, since 2020 this population is instead identified using administrative data removing and adding records in the register on the basis of life signals under the assumption of absence of under coverage of the extended Population Register (PR): the register including also the workers and the students not resident (Zindato, Bernardini, Chieppa, Cibella, Solari, 2022 for more details). The information collected by census sample surveys planned on the basis of MS design provide population structural information at high level of territorial granularity and for very detailed level of structural domains. That allows to add to the set of register variables sample estimates for variables that are not deduced from the administrative sources. These variables, such as non-employment status and commuting, must be estimated by means of sample surveys exploiting data coming from the PR as auxiliary information. In the following such type of information will be referred as survey variables. The correspondent estimates, computed mainly by means of indirect estimators, are expected to address specific requirements in terms of:

1. *accuracy*, which measures the closeness of the estimates to the true value of the target parameter;
2. *efficiency*, measured in terms of standard errors or coefficients of variation that need to be lower than prefixed thresholds;
3. *level of detail*, that is needed to satisfy the relevant information required
4. *consistency*, that assures the coherence of the estimates produced at different levels of detail.

With the aim of computing these estimates and the coverage of the PR, the MS is based on a two- phase sampling design with two different component samples, namely A and L. The component A is a sample of enumeration areas and/or of addresses selected from an Integrated Address File. It has been designed for estimating under-coverage and over-coverage rates of the PR, at national and local level, for different sub-population profiles given by different combinations of sex, age and nationality. These rates have been estimated and applied to the PR, under the form of correction weights, for obtaining weighted population counts corrected for coverage errors only. Correction weights were calculated only for the first two yearly waves (2018 and 2019) of the Permanent Census. Component L is a sample of households designed to estimate census target variables that can not be measured from registers. In order to exploit all the collected information, for the first survey cycle from 2018 to 2021 both surveys had the same questionnaire. As a consequence, sample A was pooled with sample L to increase the precision of hypercube estimates for the 2018-2021 cycle. Furthermore the component L, provides information on non-contacted people in the field and it has been used to improve the reliability of the domain estimates of PR over-coverage rates. For the second cycle (2022-2026) the questionnaires of the two surveys are different and the survey A will no longer be used to estimate PR coverage, but only to provide quality indicators on the counting process based on administrative information. More in detail, components A and L are yearly sample, whose size is about 450,000 (300.000 for the second cycle) and 950,000 households, respectively. The households are drawn from 2,850 sampled municipalities (out of about 7,950). About 1,150 municipalities are self-representative, the remaining are instead selected according to a rotation sampling scheme. Only the A survey is carried out in municipalities with size below 1,000 individuals. Municipalities with less than 300 inhabitants are instead completely surveyed. All the Italian municipalities are sampled in the four years 2018-2021 (the five years 2022-2026 for the second cycle). In the following we will focus on the estimation methods applied to compute from the MS data the required estimates of target census hypercube, both for each annual wave and with reference to the decennial outcome (which coincides with the year 2021 for the first cycle), as required by Eurostat regulations. This is a more traditional situation in which the main source of data are the MS survey data in which information from registers and others administrative data provides unit-level auxiliary information that can be used in the estimation phase to improve the quality level of the estimates to be produced. In this context, Small Area Estimation (SAE) methods are useful and needed especially when, as result of too small domain sample sizes, for some structural and/or territorial domains the coefficients of variation of the direct estimates are too high, *e.g.* larger than 0.25. Furthermore, it is also important to consider that some municipalities are not observed every year. In this situation, SAE unit-level methods are needed in order to compute estimates of the target parameters at the required level of disaggregation. During the first census cycle, with the aims of ensuring micro-macro coherence of the results the state of the working, non-working and commuting population have been predicted at unit level by means of SAE multinomial fixed effect models. Anyway, others methodologies based on small area mixed models can be taken into account: Among them, a small-area estimation method called mind (Multivariate Inference for Domains) based on multivariate and multi-effects estimation techniques has been proposed (D'Alò, M., S. Falorsi., A. Fasulo, 2022). It also provides the opportunity to introduce spatial and/or temporal correlation among random effects. These models could improve the quality of estimates for small domains of interest, can be applied for estimating parameters of interest even for out of sample domains and may result well suited in integrated frameworks, like the one under study, where the main social surveys are integrated with the population census system. About this we note that the Census and Social Survey Integrated System (CSSIS) is the third pillar of an overall framework whose aim is to obtain an integrated data - base containing, at different levels of granularity, information gathered from the census and social surveys. In Istat, a stovepipe approach has so far followed to carry out the main large-scale surveys. This has implied that for each survey, independent survey designs and different sampling strategies with different methods of collecting data in the field are involved. Therefore, it can not be taken for granted that the direct estimates of the same target parameters, produced by the different data production processes, are consistent with each other. Moreover, this approach does not enable to exploit information observed in other surveys, consequently, estimates may be less accurate than the correspondent estimates in which the available information is harmonised and used in an integrated way. To solve these problems, the goal of CSSIS is to obtain more accurate, coherent and efficient estimates, using the integrated information available. With this aim, the sampling strategies of the most important social surveys could be planned as a two phase sampling design, in which the second phase samples are drawn from MS. In particular a set of negatively coordinated samples of households are selected for the second phase surveys, aiming at providing information on the main social survey variables, such as those observed by Labour Force Survey (LFS), Aspects of Daily Life (ADL) and EUSILC. Figure 2.1 shows an overview of CSSIS framework. For each unit of PR register variables are known, while survey target variables may be unknown and needs to be collected by means of surveys (Census and/or social surveys).

The first-phase sample (*i.e.* collected with MS) information allows to estimate Census figures and provides auxiliary information for the second phase round, in which the variables of interest for the social modules are collected. Nested, non-nested or partially nested second stage samples can be drawn from MS. In the Figure 2.1, ADL module is a nested second phase sample drawn from MS, LFS and EUSILC as portrayed in the Figure are partially nested second phase samples.

Figure 2.1 - An overview of the CSSIS



In the first Census cycle (2018-2021) ADL has been already selected as module of MS: the municipalities are selected among the MS sampled municipalities that are stratified at the regional level with respect to their population size; the households for ADL are drawn among the households of the MS. Therefore, MS and ADL represent a classical nested two phase sampling strategy. The methodological complexity of this design is given by the different stratifications that can be used in the two surveys and for the use of stratified two-stage sampling design for both surveys. At the moment, LFS and EUSILC are still independent surveys even if, for LFS a sub-sample of the municipalities has been selected for the MS. The implementation of an integrated sampling strategy for LFS and EUSILC should also take account of the additional complexity caused by their household rotation group scheme. It still under study the best way of integrating these surveys with the Census MS, given the estimation goals of each survey and the constrain of minimising the statistical burden.

From a general point of view, an integrated sampling strategy can allow to generate an Integrated System of Microdata in which several different blocks, defined by subsets of units and available variables, can be identified and properly treated in the estimation phase.

- The first block is defined considering all units in the PR and the register variable (green block and yellow block too, after proper statistical treatments, in Figure 2.1). In this case, aggregated values are obtained just summing up all the information at a required level of granularity.
- The second block is defined considering the intersection between MS and PR (red block in Figure 2.1). It is given by a subset of units belonging to PR for which, besides the register information, also survey variables collected by MS are available. In this case, register variables can be used in the estimation phase as auxiliary information and target variables collected by MS can be estimated via calibration methods or model-based small-area estimators.
- Others blocks can be defined considering the intersection between MS and PR and each of second phase samples (modules) drawn from the MS (blue blocks in Figure 2.1). Each of them is given by a subset of units belonging to PR for which besides the register information also survey variables collected by MS and by each single module of social survey are available. Also in this case, estimating via calibration methods or model-based small-area estimators for target variables collected by MS and each single module can be used. Moreover, it possible to use design and model-based projection estimators or estimators based on micro and macro integration (Kim and Shao, 2021).
- Whenever there is an intersection between the social modules, further blocks can be defined taking into account the intersection between PR, MS and each pair of modules. In particular, information from two or more second-phase blocks can be jointly exploited, *i.e.* integrated, in order to obtain design-based or model-based estimates that are more efficient than the corresponding estimates in which the blocks are not exploited in an integrated way.

The estimates of interest of each table, derived from the above blocks, can be computed by means of design-based method, as long as the sample is large enough to yield reliable estimates. When the sample size associated with a very detailed table is not sufficiently large, small area methods need to be applied, investigating how these methods can be used in this framework and how the consistency between tables computed at very detailed level by model-based small area estimators and marginal tables computed by direct estimators can be pursued. Integrating the available information and by modelling the relationship among targets variables and a set of covariates, efficient and reliable estimates can be obtained for instance by means of projection type estimators. This set of methods aims to improve the efficiency of the estimates under consideration, using unit-level auxiliary information of a small sample, *i.e.* the second phase sample, also available in MS or PR. The working model is fitted by using data from a survey in which the target variable is observed. This model is then used to calculate predicted values of the target variable either in the bigger survey, MS, or stored in a sort of augmented register, say augmented PR, that includes for each unit besides the register variables also the observed/predicted values of survey target variables. When design-based method can be applied, estimation methods can be defined under a model-assisted framework, so that the inference does not depend on the validity of the working model. The gain of efficiency increases as the correlation between target variable and auxiliary information increases. The method's advantage lies on the fact that when predicted values of target variables are generated, each final estimate can be easily obtained, by using a unique set of sampling weights associated to the records collected through MS, or simply by summing up the corresponding values in the PR, when the prediction is performed on the register.

3. A generalised estimation method for CSSIS

As stated before, the estimates of each table of interest, derived from the above blocks, can be computed by means of design-based method as long as the sample is sufficiently large to yield reliable estimates and all target domains are covered by the sample. Small area methods should be applied when the sample size associated with a very detailed table is not sufficiently large to produce reliable direct estimates or some domains are out of sample. In this case, one of the most important research topics for the census in the coming years would be the improvement of the small area estimation methods that can be applied, by studying, for example, multivariate models that could also exploit spatial and/or temporal correlation of the residuals. With the increasing of Census and CSSIS surveys waves, temporal sample information can be pooled and SAE models exploiting temporal correlation would became more effective in improving the level of quality of SAE estimates.

A multi-effect estimation method proposed and developed with mind (D'Alò, Falorsi, Fasulo, 2022) could be address this issue. This method is a multivariate version of the standard a small area estimators based on General linear mixed model carried out following the approach of (Datta, Day, Basawa, 1999). It can allow to introduce further marginal random effects, in addition to the usual single area random effect, that can be useful to better keep under control the potential bias due to the synthetic part of the model, whenever the areas of interest are very small or when some of them are out of sample. The marginal effects may include some of the design variables used to define strata or to define planned domains.

With reference to the introduction of multivariate models, it is important to note that in the Census framework, a precise multivariate model specification can coincide with a full tabulation plan, in which multiple contingency tables made up crossing numerous territorial and structural classification variables need to be estimates. In this framework, multivariate and multilevel model-based estimation methods could improve the efficiency of the results with respect to those that can be obtained using standard univariate small area estimation approach. In fact, from one side, a unique model specification can affect the efficiency level of each variable, since predictive power of the model varies as the target variable changes; on the other hand, a unique model can guarantee in a simple way a greater level of overall coherence among the different sets of estimates, than that obtainable by setting a specific model for each target variable. Obviously, the model level specification, as well as the reference levels for fixed and random effects, has an impact on the accuracy and efficiency of the entire set of produced estimates.

References

Boeschoten, L., D. Filipponi, and R. Varriale. 2021. "Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status". *Journal of Survey Statistics and Methodology*, Volume 9, Issue 3: 549-573.

D'Alò, M., S. Falorsi., and A. Fasulo. 2022. "MIND, a methodology for multivariate small area estimation with multiple random effects". Presentation at the *7th Italian Conference on Survey Methodology* - ITACOSM 2022. Perugia & Assisi, Italy, 7th - 10th June 2022.

Datta, G.S., B. Day, and I. Basawa. 1999. "Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation". *Journal of Statistical Planning and Inference*, Volume 75, Issue 2: 269-279.

Filipponi, D., U. Guarnera, and R. Varriale. 2019. "Hidden Markov Models to estimate Italian employment status". Presentation at *New Techniques and Technologies for Statistics* - NTTS 2019. Brussels, Belgium, 11th – 15th March 2019.

Ioannidis, E., T. Merkouris, L.-C. Zhang, M. Karlberg, M. Petrakos, F. Reis, and P. Stavropoulos. 2016. "On a Modular Approach to the Design of Integrated Social Surveys". *Journal of Official Statistics - JOS*, Volume 32, Issue 2: 259–286.

Kim, J.K., and J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach". *Biometrika*, Volume 99, Issue 1: 85-100.

Kim, J.K., and J. Shao. 2021. *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL, U.S.: Chapman and Hall/CRC.

Office for National Statistics - ONS. 2016. "Annual assessment of ONS's progress towards an Administrative Data Census post-2021". *Data and Analysis from Census 2021*. Newport, South Wales, UK: ONS (Accessed on 23rd February 23 2023). https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs.

Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture". *Journal of Survey Statistics and Methodology*, Volume 3, Issue 4: 425-483.

Zindato, D., A. Bernardini, A. Chieppa, N. Cibella, and F. Solari. 2022. "Estimation of population counts: lesson learned from the first cycle of the Italian Permanent Population Census". Presentation at the *7th Italian Conference on Survey Methodology - ITACOSM 2022*. Perugia & Assisi, Italy, *7th* - *10th* June 2022.

Multi-source statistics in the Italian permanent census

Marco Di Zio[1], Danila Filipponi[1]

Abstract

The Italian Census of Population and Housing is based on the integration of sample survey data and administrative information. The statistical procedures for the prediction of attained level of education and the employment status are representative of two approaches when dealing with multi-source data. The attained level of education is set in a context in which one of the sources can be taken as reference for the target variable under study (supervised approach), i.e. the target variable is assumed free of errors. For the emplyment status, all the available data sources are assumed to be affected by errors and therefore the prediction is modelled as a latent variable computed conditionally on the observed values of the data sources (unsupervised approach). The paper describes the two procedures, and discusses differences and common aspects, as well as relevant points to take into account in the register-based multi-source estimation of attained level of education and employment status.

Keywords: Register-based statistics, mass imputation, data integration, latent model, log-linear imputation.

1. Introduction

Since October 2018, the Italian Statistical Institute is conducting yearly, during the first week of October, the Census of Population and Housing on a sampling basis in order to collect updated information on the main characteristics of Italian resident population and its social and economic conditions at national, regional and local levels. The project consists on the production of data for the entire population through the integration of census sample survey data, the base register of individuals (BRI) and administrative information. The sample surveys are conducted for dealing with errors and lack of information in the administrative sources. In accordance with

[1] Marco Di Zio (dizio@istat.it); Danila Filipponi (dafilipp@istat.it), Italian National Institute of Statistics - Istat.

this approach, descriptive statistics can be derived by directly computing the indicator of interest using unit-level data estimated for the entire population. This approach is in line with the rising demand for more granular and comprehensive statistics. In this respect, it is important to underline differences among the Census target variables. Variables as gender, place and date of birth, citizenship are obtained by integrating only administrative data, and can be thought free of errors or, more honestly, affected by a negligible error. Other core variables, like the *attained level of education* (ALE[2]) and the *employment status* (OCC) are estimated by integrating administrative data and sample surveys. For their construction, statistical models are used and consequently they are affected by a natural degree of uncertainty that should be taken into account in their usage.

The statistical procedures for the prediction of ALE and OCC are representative of two approaches when dealing with multi-source data. The approach for the ALE estimation is set in a context in which one of the data sources can be taken as a reference for the target variable (*supervised approach*), *i.e.* target variable is assumed free of errors. On the other side, for OCC, all the available data sources are assumed to be affected by an error and therefore the prediction is modelled as a latent variable computed conditionally on the observed values of the data sources (*unsupervised approach*). More in detail, the methods used for the reconstruction of the ALE are log-linear models, while a mixture of latent Markov models are applied for employment.

Despite those differences, there is a common aspect in the application of these two procedures in the Italian permanent census, they aim at estimating a value of the variables for each unit in the population register through a random draw from the estimated probability distribution. This approach, which naturally increases the variability of the data and estimates, has the advantage of better representing the probability distribution of the variable, thus ensuring greater flexibility in their usage. On the other hand, this advantage may transform to a risk, because users can be tempted to use microdata without any limitation. For this reason, it is of fundamental importance to provide a flexible tool for measuring uncertainty of estimates at various unplanned level of aggregation.

[2] ALE classification: 1 – Illiterate, 2 - Literate but no formal educational attainment, 3 - Primary education, 4 - Lower secondary education, 5 - Upper secondary education, 6 - Bachelor’s degree or equivalent level, 7 - Master’s degree or equivalent level, 8 - PhD level.

2. Multi-source supervised and unsupervised modeling

2.1 Estimation of the attained level of education

The estimation of the attained level of education integrates administrative data sources from the Ministry of Education, University and Research (MUR), the 2011 Italian census and sample surveys. While units of the sample surveys are spread all over the population, the other data sources are related to specific sub-populations. With some simplifications, the following subgroups can be defined:

a. Subgroup A is composed of all persons with administrative information on ALE at time *t*-2. It is characterised with longitudinal information on course attendance for people entering a study programme after 2011 to *t*-2, scholar year (*t*-2,*t*-1), approximately 22% of the population of people older than 9 years.

b. Subgroup B is composed of persons not enrolled in any school course included in administrative data from 2011 to *t*-2, with information from 2011 Census, approximately 73% of the population of people older than 9 years.

c. Subgroup C is composed of individuals neither in MIUR nor in 2011 Census. For this group, no direct information on ALE is available. This subgroup is composed mainly of adults and is characterised by a high percentage of not Italian people, approximately 5% of the population older than 9 years.

It is important to remark that administrative information has some coverage errors: It does not include some particular qualifications, like *Fine Arts, Drama, Dance and Music* academic diplomas and some courses managed by regions. An additional critical issue is due to the time lag between the availabily of administrative data and the reference time. Also 2011 census has some educational levels of the actual classification not included in their data.

The lack of joint information of education level referring to the same reference period, and of some classifications, motivated the use of a supervised approach.

The procedure adopted is based on log-linear imputation (Di Zio *et al.* 2019). In particular, we estimate the conditional probabilities of ALE at time *t* (ALE_t) given a set of covariates *X*, $Pr(ALE_t \,|X)$, then we impute ALE$_t$ by randomly taking a value from this distribution. The conditional probabilities $Pr(ALE_t\ |X)$ are estimated as follows: First, a log-linear model is applied to the contingency table obtained by cross-classifying the variables (ALE_t, X) to estimate their expected counts $\hat{\eta}_{ij}^{ALE_t\ X}$, from which the counts $\hat{\eta}_j^X$ are derived. The estimated conditional probability distribution $\widehat{Pr}(ALE_t|x)$ is easily obtained by computing $\hat{\eta}_{ij}^{ALE_t\ X}\ /\hat{\eta}_j^X$.

As stated by Singh (1988), this method generalises hot-deck imputation by choosing suitable predictors for forming “optimal” imputation classes. The approach is based on modelling the associations between variables. It includes as a special case the random hot deck when all the interactions between variables are included in the model (saturated log-linear model), but has the advantage of allowing the use of more parsimonious models by testing the associations among variables. This is an important characteristic especially when the number of variables and contingency table’s cells increase.

In subgroup A, given the great informative capacity of administrative data, we estimated $Pr(ALE_t\ |X)$ by using only administrative data. Information on ALE in the year *t-2* and information on year attendance of educational courses in academic year (*t-2, t-1*) are available for all the units.

Administrative data allow estimating the probability of obtaining a new qualification based on schooling characteristics in 2 years, *i.e.* $Pr(ALE_{t-2}|$ $ALE_{t-4,}$ *age, citizenship, school attendance*). This probability is used to predict $Pr(ALE_t|$ ALE_{t-2}, *age, citizenship, school attendance*). The assumption is that probabilities are stable in a short period like that of this application.

For the other two subgroups, we estimated $Pr(ALE_t\ |X)$ by using ALE_t observed in the sample as a target variable.

More specifically, log-linear models for subpopulations B and C are built to estimate the following conditional probabilities:

a. Subgroup B: $Pr(ALE_t|$ ALE_{t-2}, *age, citizenship, province of residence, gender);*

b. Subgroup C: $Pr(ALE_t\ |$ *age, citizenship, gender, apr, sirea).*

Apr is an auxiliary information on ALE coming from an administrative source and it covers a particular subpopulation of individuals: Those who changed their place of residence after 2014. It is a self-declared ALE and it comes with a more aggregate classification (4 levels[3]). *Sirea* refers to people who were targeted but not surveyed by the 2011 Census and were later detected by post-Census operations carried out in agreement with Italian Municipalities.

A model selection step based on cross-validation imputation is carried out, it is important to notice that most of the times imputation is based on saturated log-linear models.

2.2 Estimation of the employment status

The validity of any supervised approach relies on the important requirement that at least one data source provides a correct measure of the target variable. Generally, survey data is treated as the privileged source of information, while the other data sources play essentially the role of covariates within a prediction approach.

A different approach is based on the assumption that all the available sources could be potentially affected by measurement errors and a possible remedy to overcome the deficiencies of the available sources is to treat them as multiple measures of the true target variable which is assumed to be unmeasurable. A prediction of the true values can then be obtained using latent variable models. Some recent examples on linked surveys and administrative data that address the problem of measurement errors with a latent variable models are given by (Boeschoten *et al.* 2019, Boeschoten *et al.* 2020, Oberski *et al.* 2017, Guarnera *et al.* 2016, and Pavlopoulos *et al.* 2015).

In order to predict the employment status for the Italian population over 15 years old, the statistical information collected during the Census operation are integrated at unit level with other two sources of information on employment: (i) the Labour Force Survey (LFS) and (ii) administrative data.

LFS is, of course, the main European survey to produce quarterly estimates on the employment status. Then, despite sampling errors and deficiencies in the survey response process, LFS is the key survey to measure correctly the employment status.

[3] Apr 4 levels of classification: 1 - Up to primary education; 2 - Lower secondary education; 3 - Secondary and short cycle tertiary education; 4 - Tertiary and post tertiary education.

Administrative data relevant for the labour statistics come primarily from social security and fiscal authorities. After an harmonisation process, data are integrated and organised in an information system having a linked employer-employees structure; from this structure it is possible to obtain information on administrative employment status for the complete population and for every month of the reference year. Nevertheless, the definition of employment in the administrative sources is quite different for the one in the statistical surveys since it depends on the administrative definition. The main errors of administrative sources on employment concern the lack of coverage of irregular workers and, for some of the sources, delays or temporal misalignments in the communications of job positions.

Census survey collects information on the employment status related to the first week of October.

Even if the definition of employment status in the European Census regulations coincides with the ILO definition and therefore with that of the LFS survey, its measurement is definitely less precise. This is mainly due by the survey questionnaire and the time interval that may occur between the interview and the reference period of the survey.

Both micro and macro comparisons of data show a level of discrepancy that cannot be overlooked, with clear indications that, despite the different levels of accuracy, none of the sources can be considered as error free, and taken as a benchmark in the estimates. However, although administrative and survey data are both affected by measurement errors, these not only do not coincide, but are complementary in identifying the target measure of employment. In fact, on the one hand, survey data make it possible to capture the population not covered by the administrative sources and on the other the administrative data allow to correct the under-reporting associated with the survey response process. So, we have multiple measures that attempt to get the unmeasured employment status.

Here, the *true* employment status at time t for subject k is modelled as a binary latent variable $L_{k,t}$ taking values 0 or 1 depending on whether at time t subject k is employed or not, respectively. The process L is analysed at the finite collection of times $t = 1, \dots, T$ and $L_{1:T}$ denotes the random vector $L_1, \dots, L_T$. In this work $T = 12$ and each time t corresponds to a specific month of the year. We are interested in the employment estimate for the month of October.

Information from Census survey, LFS and administrative sources are treated as imperfect measures of the target process. Specifically, $Y_{1:T}^{i}$ with $i = 1,2$ denote the binary vectors of (possibly missing according to the sampling design) values of the employment status at times $1, \dots T$ resulting from the two surveys, while a third measure is a dichotomous vector $Y_{1:T}^{3}$ whose components are 1 (employed) for a certain individual if he, or she, appears in at least one of the original administrative sources, and 0 otherwise. Individual covariates X like *gender*, *age* class (5 levels), *income* class (5 levels), *level of education* and two binary flags associated with retirement status and being a student are used to explain different behavioural characteristics in the employment dynamics. Moreover, a four-category covariate S is defined to account for different quality levels of the different typologies of administrative sources.

The model specification requires the definition of two parts: The latent model which describes the distribution of the latent variables and the manifest model which describes the conditional distribution of the observed variables given the latent variables. The latent component is specified to be a mixture of Markov model. Specifically, heterogeneity in the employment activities are explicitly modelled through a categorical latent variable G with three components corresponding to three different sub-populations: Never working people, individuals with stable employment dynamics and people who are likely to change frequently their employment status. Since the characteristics of the three groups are likely to be strictly related to demographic features as well as to the type of administrative source information is taken from, the distribution of the latent random variable G is modelled conditional on covariates X and S. In the following, $\Phi_{g|s,x}$ will denote the conditional probability, where $g = 1,2,3$ and (x, s) are realisations from the variables X e S. The employment dynamics L for each sub-population g is governed by a first order Markov chain with initial probabilities $\tau_{i}^{g} = P(L_{i} = j|G = g)$ and transition matrix M^{g} whose typical element $\{M_{jk}^{g}\}$ will be denoted by $\pi_{k/j}^{g} = P(L_{t} = k|L_{t-1} = j, G = g),\ j, k = 0,1$. The assumption of a Markov Chain can be a valid assumption for the sub-population $G = 2,3$, whereas is unrealistic for the sub-population of never working people. We can assume that the latent process L for $G = 1$, is a degenerate Markov chain.

The next step is the definition of the measurement model, *i.e.* the probability distribution of the random vectors $Y_{1:T}^{i}$ with $i = 1,2,3$ given the latent process and the covariates. According to a common approach, we assume that, conditionally on the latent process, the three measurement processes are independent one of each other and that measures associated with LFS, admin sources, and Census at time t are independent with the corresponding measures at different times (serial conditional independence). Moreover the manifest variables related to the administrative data are supposed to depend on the covariate that describe the type of administrative information S, while no covariates are introduced in the specification of the distribution of the surveys data. Due to the conditional independence assumptions, the relevant parameters of the observational model are $\psi^{g}_{j|i} = P(Y^{g}_{t} = j|L_t = i)$ for $g = 1,2$ and $\psi^{3}_{j|i,s} = P(Y^{3}_{t} = j|L_t = i, S = s)$ for the administrative sources, with $= 1, \dots, 12,\ (i,j) = \{0,1\},\ s = \{1,2\ 3\ 4\}$. An important constrained introduce into the model is $\psi^{1}_{0|1} = 0$ which means that "no false positive" data are present in the LFS data. This constraint relies on the empirical evidence that unemployed people are unlikely to declare they work. This hypothesis fails in Census due to time shifting of the responses with respect to the actual working period.

Parameters are estimated by maximising the log-likelihood which can be done by EM algorithm. Then, the marginal conditional distributions of the latent variables given a manifest configuration can be obtained by Bayes' Theorem and the employment status for each month and each configuration can be scored by generating from them.

3. Validity and accuracy of multi-source supervised and unsupervised modeling

The assessment of the procedure and results is a crucial point, especially for these complex settings that involve data of different nature and statistical models.

In the supervised approach, observed data can be used for computing the usual goodness-of-fit measures, and register-based estimates computed on BRI data can be compared with those of the sample survey. Since the objective of the procedure is essentially that of producing a pseudo-population, *i.e.* data that can be thought as generated from the probability distribution of the target variable, the evaluation is carried out by comparing the probability distribution obtained by means of BRI with those computed on the data of the sample (see Di Zio *et al.* 2019).

Another important assessment is that pertaining the data validation from a content perspective. Subject matter experts have analysed and compared the results with other information related to the ALE. This is a particularly important step for the production of statistics in a National Statistical Institute.

Finally, a measurement of accuracy of estimates should be produced. This is an open methodological issue, because of the fact that it should include all the uncertainty sources, *e.g.* model uncertainty, sampling error, coverage errors, and so on. The problem is still under investigation when dealing with register-based statistics. Some references are (Scholtus *et al.* 2021, Alleva *et al.* 2021). For the first ALE applications, a replication approach was carried out to have an idea of the error. The procedure replicates the sample and imputation procedure, but with a stratified sampling design that is a simpler schema than that used in practice essentially based on a two stage cluster sampling. Replication approaches are appealing, especially in a context complex in terms of error source and model applications, but they are computationally demanding and given the dimension of the problem are difficult to apply in the production line.

In (Di Zio *et al.* 2022), there is a study for an analytical approach to the variance evaluation of estimates. Reminding that most of the predictions are obtained through saturated log-linear imputations, they resort to the application of classic formula for variance estimation with random hot deck within imputation classes defined by the auxiliary variables chosen for each segment of the population.

As far as the variance of estimates of population B and C, $V\left(\hat{\bar{Y}}_B\right)$ and $\mathrm{V}\left(\hat{\bar{Y}}_C\right)$ are concerned, we may adapt the basic formula for the variance estimation in presence of donor imputation (see Wolter 2007, appendix F2, Brick *et al.* 2004) obtainin

$$V\left(\hat{\bar{Y}}_B\right) = \frac{\left(1-\frac{n_B}{N_B}\right)\sigma^2_{y_B}}{n_B} + \frac{1}{(N_B)^2}\sum_k \left(N_{x_k} - n_{x_k}\right)\left(1-\frac{1}{n_{x_k}}\right)\sigma^2_{y_B|x_k}$$

where x_k for k=1,…, K are the imputation cells, N_B is the population size of B and n_B is the size of the sample s falling in B, $\sigma^2_{y_B}$ is the variance of Y (ALE) in the population B, and $\sigma^2_{y_B|x_k}$ is the variance of Y in B within stratum $x_{k.}$ An analogous formula can be derived for subgroup C. If the auxiliary variable

X is strongly connected to *Y*, an estimate for $V\left(\hat{\hat{Y}}_B\right)$ can be obtained by using the sampling variance of *y* within stratum x_k, $\sigma^2_{y_B|x_k}$, for both terms. In the first term the conditional variance should be obtained by a weighted sum of conditional variances with weights given by the square of the size of the strata.

For the subgroup A, a slightly different formula should be derived. We remind that in this subset the predictions are obtained by estimating a log-linear model on previous data, and by applying the estimated model to the actual data

$$\hat{V}\left(\hat{\hat{Y}}_{CNG}\right) = \frac{\left(\frac{N_A - n_A}{N_A}\right)^2 \left(1 - \frac{N_A - n_A}{N_A}\right) \hat{\sigma}^2_{ADM_A}}{N_A - n_A} + \frac{\left(\frac{n_A}{N_A}\right)^2 \left(1 - \frac{n_A}{N_A}\right) \hat{\sigma}^2_{s_A}}{n_A}$$

This formula is similar to the previous one, but $\hat{\sigma}^2_{ADM_A}$ is estimated only on administrative data at time *t-2* without resorting to the sample *s*, while $\hat{\sigma}^2_{s_A}$ is estimated by using units of sample *s* that are in A.

The validation of estimates obtained for the employment status according to a latent variable approach is less standard and more complex. Different latent class models have been estimated and the model that better fit the observed data has been assessed based on model criteria like Akaike and/or Bayesian information criterion. Together with the classical model-based criteria, the final selection among candidate models have been carried out also considering the model interpretability. Indeed the identification of the optimal model is not always clear and the validity of the approach is mainly demanded to sector experts who have the task of verifying that the latent construct identified by the model effectively corresponds to the target variable.

It is necessary to emphasise that unlike the latent variable models used in psychometrics where a difficult process is required for the identification of a latent construct, which includes the identification and validation of the elements associated with the latent construct, when the latent models are used to deal with measurement errors, the identification and the evaluation of the latent variable model is simpler. One major difference is that the definition of latent variable *L* and the number of classes it includes is not a research question, but they are known a-priori and must be the same as the number of levels observed in the manifest indicators. The primary reason for introducing multiple indicators of *L* is to improve its measurement. A well-defined and measurable construct with a known number of classes is rather crucial, otherwise *L* cannot be regarded as the true characteristic. In this contest latent class response probabilities $\psi_{j|i}^{g}$ can be identified as error and used for estimating the classification error for one or more of the indicators.

Also under this scenario a measurement of accuracy of estimates should be produced. Boeschoten *et al.* 2020 developed a multiple imputation procedure denominated MILC. The mixture of latent Markov models is used to impute the latent construct under evaluation and the parameter uncertainty is dealt with within a frequentist framework by using a nonparametric bootstrap. The MILC procedure comprises five steps. In the first step, *m* nonparametric bootstrap samples are selected from the observed frequency distribution. In the second step, the mixture latent Markov model is fitted on each of the *m* bootstrap samples. Then in the third step, *m* prediction of *L* are created using the *m* estimated parameters obtained from the *m* bootstrap sample. These imputations can be created using either the conditional imputation procedure or the marginal imputation procedure. In the fourth step, estimates of interest can then be obtained from every imputation, and in the fifth step, the estimates obtained for every imputation can be pooled using the pooling rules defined by (Rubin 1987).

A monte carlo study is implemented to evaluate the performance of the imputation procedures and to investigate whether the MI is an appropriate method to evaluate the variability when a latent Markov model is used to impute a latent construct. The simulation highlighted the usability of the MILC method in different conditions. A limitation of the current simulation study is that classification error rates larger than 20 percent were not investigated (Boeschoten *et al.* 2020).

A multiple imputation approach is interesting and so far it seems the only solution even if is computationally demanding given the dimension of the problem.

Although the methods are separately described, ALE and OCC are potentially dependent, thus it is worthwhile to remark that, in the census application, the model for OCC includes ALE as an explicative variable, while ALE is estimated independently of OCC.

4. Final remarks

This paper presents two multi-source estimation approaches adopted in the Italian permanent population census. They relies on different assumptions and represents two strategies of using multi-source data. One is supervised and the others is unsupervised. The supervised approach is mainly adopted because observed data cannot be considered a multiple measurement of the target variable ALE. In the second, information can be more likely interpreted as a multiple measurement affected by errors of the target variable OCC.

In such a complex setting, an important task is that of assessing the quality of estimates obtained with predicted data. Evaluations are essentially based on the usual goodness-of-fit measure according with the adopted model, evaluation by subject experts - which are of particular importance for NSIs - and accuracy measures for register-based estimates. On the latter issue, some experiences are reported. Since the aim is that of providing a set of microdata on which a user may easily compute estimates and their accuracy, a flexible tool for computing accuracy is desirable. For ALE, it is adopted an analytical approximation, while multiple imputation is considered for OCC. The analytical approach is certainly a useful tool, but it is strictly connected to the method used that is a sort of stratified random hot deck, moreover it resorts to some approximations that may fail when domains of estimates become small. Multiple imputation can be a reference tool because it is in principle designed with the aim of allowing the user to evaluate uncertainty of unplanned estimate through the release of multiple microdata. Nevertheless, its use in a NSI context is still a problem especially in terms of managing and even more of accepting the idea of having ‘more’ potential registers (multiple registers).

References

Alleva, G., P.D. Falorsi, F. Petrarca, and P. Righi. 2021. “Measuring the Accuracy of Aggregates Computed from a Statistical Register”. *Journal of Official Statistics - JOS*, Volume 37, Issue 2: 481-503.

Boeschoten, L., T. de Waal, and J.K. Vermunt. 2019. "Estimating the Number of Serious Road Injuries Per Vehicle Type in the Netherlands by using Multiple Imputation of Latent Classes". *Journal of the Royal Statistical Society, Series A: Statistics in Society,* Volume 182, Issue 4: 1463-1486.

Boeschoten, L., D. Filipponi, and R. Varriale. 2021. "Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status". *Journal of Survey Statistics and Methodology*, Volume 9, Issue 3: 549-573.

Brick, J.M., G. Kalton, and J.K. Kim. 2004. "Variance Estimation with Hot Deck Imputation Using a Model". *Survey Methodology*, Volume 30, Issue 1: 57-66.

Di Zio, M., R. Filippini, and G. Rocchetti. 2019. "An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data". *Rivista di statistica ufficiale/Review of official statistics*, N. 2-3/2019: 143-174. Roma, Italy: Istat. https://www.istat.it/en/archivio/271219.

Di Zio, M., R. Filippini, and S. Toti. 2022. "Variance estimation for the mass imputation of the "Attained level of education" in the Italian Base Register of individuals: A comparison between analytical and MonteCarlo estimates". Paper presented at the *Expert Meeting on Statistical Data Editing*. 3rd-7th October 2022 (virtual), United Nations Economic Commission For Europe Conference of European Statisticians. https://unece.org/statistics/events/SDE2022.

Guarnera, U., and R. Varriale. 2016. "Estimation from Contaminated Multi-Source Data Based on Latent Class Models". *Statistical Journal of the IAOS*, Volume 32, Issue 4: 537-544.

Oberski, D., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models". *Journal of the American Statistical Association*, Volume 112, Issue 520: 1477-1489.

Pavlopoulos, D., and J.K. Vermunt. 2015. "Measuring temporary employment. Do survey or register data tell the truth?". *Survey Methodology*, Volume 41, Issue 1: 197–214.

Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*. Hoboken, NJ, U.S.: John Wiley & Sons, *Wiley Series in Probability and Statistics*.

Scholtus, S., and J. Daalmans. 2021. “Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data”. *Journal of Official Statistics - JOS*, Volume 37, Issue 2: 433-459.

Singh, A.C. 1988. “Log-linear imputation”. *Working Paper, Methodology Branch*, N. SSMD-88-029 E. Ottawa, Canada: Statistics Canada.

Wolter, K.M. 2007. *Introduction to Variance Estimation*. New York, NY, U.S.: Springer.

SESSION 2

Methodologies for multi-source processes

INTRODUCTION

Natalie Shlomo[1]

National Statistical Institutes (NSIs) are assessing the availability of new sources of data to complement and improve statistical processes. With continuing decline in response rates and high costs of conducting large-scale surveys and censuses, more focus has been directed to the research and development of producing multi-source statistics. There have been several European initiatives to promote this research, particularly the ESSnet project on the Quality of Multi-source Statistics (KOMUSO)[2]. Open issues for research around multi-source statistics includes scoping out new data sources and how they can be integrated into the statistical processes at NSIs. Traditionally, these data sources were based on administrative data[3] but more attention has also been directed towards big data sources[4].

Some of our standard methodologies in the official statistics tool-kit for producing multi-source statistics need to be revised and enhanced. Data integration techniques and record linkage need to be improved to account for the case where there are limiting matching variables between data sources. More focus is needed on improving statistical matching and mass imputation techniques for combining data sources and producing statistical registers that can also accommodate small area estimation. Other standard methodologies, such as compensating for missing data, estimation and calibration, and accounting for measurement errors, all need special consideration when used in the context of combining data sources to produce multi-source statistics. Moreover, particularly challenging for producing multi-source statistics is the need to improve existing quality frameworks that account for data that is 'ingested' into the NSI from other government or commercial organisations and hence not specifically collected for statistical purposes. Closer relationships and agreed quality frameworks are needed across the organisations providing data for statistical purposes within the NSI.

[1] Natalie Shlomo (natalie.shlomo@manchester.ac.uk), University of Manchester.

[2] See: https://ec.europa.eu/eurostat/cros/content/essnet-quality-multi-source-statistics-komuso_en.

[3] See: https://ec.europa.eu/eurostat/cros/content/admindata-essnet-use-administrative-and-accounts-data-business-statistics_en.

[4] See: https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en and https://ec.europa.eu/eurostat/cros/essnet-big-data-2_en.

Producing statistics from multiple data sources also requires new IT technologies and standards to promote an integrated eco-system that can utilises all existing data sources. This area of research has also led to advancing the use of machine learning and AI approaches in the production of official statistics. The AI approach of neural networks is demonstrated in the second presentation of this session to impute an education variable in the Register of Individuals in Italy. Using machine learning and AI approaches brings specific challenges on how to assess quality and account for uncertainty, particularly as these approaches can be seen as ‘black boxes’. These approaches have a new set of quality measures that need to be explained in our quality frameworks, such as: prediction power, F1 statistic, C statistic (area under the ROC curve) and cross-validation. It is important that research and development continue to advance the quality assessment of machine learning and AI approaches and how they are explained to the users of official statistics. One such example in this direction is the paper by Yung *et al.* (2022).

Another consideration in producing multi-source statistics is protecting the confidentiality of data subjects. The dissemination of official statistics outputs based on multiple data sources brings particular challenges with respect to avoiding breaches of confidentiality. The problem is more complex in this setting since the data custodians of the different data sources that have been ingested into the NSI have the knowledge to re-identify individuals in the statistical output.

All these challenges mentioned in this introduction should be at the forefront of research and development on multi-source statistics. The ESSnet programmes are a good way forward to promote cooperation across NSIs to collaborate on research towards the common goal of evolving our official statistics production to include multiple data sources. More programmes should be added to meet these specific challenges.

At Session 2 of the first Istat Workshop on Methodologies for Official Statistics titled: Methodologies for Multi-source Processes, the following presentations were given:

- “Overview of the Istat activities and open problems” – presented by Marco Di Zio;

- "A study of MLP for the Imputation of the "Attained Level of Education" in Base Register of Individuals" – presented by Romina Filippini.

The discussion was lead by Thomas Burg (Statistik Austria) and the session terminated with the point of view of the Statistical Production Department of Istat by Carlo Maria De Gregorio (Istat), head of the Division for Development and Enhancement of the Integrated Registers System by Theme.

References

Yung, W., S.-M. Tam, B. Buelens, H. Chipman, F. Dumpert, G. Ascari, F. Rocci, J. Burger, and I. Choi. 2022. "A quality framework for statistical algorithms". *Statistical Journal of the IAOS,* Volume 38, Issue 1: 291-308.

Overview of the Istat activities and open problems

Marco Di Zio[1], Stefano Falorsi[1], Silvia Loriga[1]

Abstract

In 2016, a deep methodological review of the Istat official statistical production system began, which moved from a traditional system to a new innovative functional scheme based on an integrated and multi-source approach. This paper starts with a description of the Istat new production process. Then, the focus is on two very relevant aspects. The first concerns the problem of choosing between two data sources, namely a sample survey and a non-probabilistic subset of the units belonging to the population. The second regards the quality of the estimates that are produced; this issue is treated by deepening two aspects: the assessment of the uncertainty of register-based statistics produced by multi-source data and the strategies to improve the coherence of the estimates. Finally, the main lines of research for the next few years in Istat are mentioned.

Keywords: Integration, Multi-source estimation, Measurement errors, Coverage errors, Non-probability samples.

1. Introduction

In 2016, a deep methodological review of the Istat official statistical production system began, moving from a traditional system to an innovative functional scheme based on an integrated and multi-source approach. It was an extremely complex innovation, which required huge investments to manage the profound changes at the organisational, operational, methodological, technological and regulatory level (in particular, for the compliance with the GDPR legislation on privacy in the new context of integration and processing of data from different sources).

The new methodological framework designed and implemented is based on the integration of three fundamental components. The first is the Integrated

[1] Marco Di Zio (dizio@istat.it); Stefano Falorsi (stfalors@istat.it); Silvia Loriga (siloriga@istat.it), Italian National Institute of Statistics - Istat.

System of Registers (ISR), which is the main methodological, technological and information infrastructure of the framework, based on massive integration processes at the micro level of data mainly coming from administrative sources, and which is itself an exclusive source of estimates for official dissemination. The second is the Permanent Census System (PCS), which provides fundamental information on the structure of the corresponding target populations (demographic, agricultural, etc.), guaranteeing very high levels of territorial and sectoral granularity and also correcting the corresponding population registers for over and under coverage. The third component is constituted by the surveys, designed to estimate variables that cannot be deduced from the other two informative systems. In particular, as regards the social surveys an innovation process was initiated, concerning the integration of the main social surveys with the Population Census in the context of the Census and Social Survey Integrated System (CSSIS).

With this new production framework, the joint exploitation of the different data sources takes place within a process that uses in addition to traditional model-assisted estimators also indirect model-based estimators for the production of micro databases through small area estimators and mass imputation techniques.

Although Istat is one of the leading countries in the application of this new production method, there is still a long way to go to reach a stabilisation of the methodological foundation of the production process. The paper, therefore, after having described the production process in more detail (Section 2), deals briefly with some of the most relevant aspects. The first aspect concerns the problem of choosing between two data sources (Section 3): the first consists of a sample survey possibly characterised by a certain non-response rate; the second source is represented by a non-probabilistic subset of the units belonging to the population of interest for which the observation of the variable of interest is available. The choice in defining one of these two sources as the main source for the production process is traditionally based prior knowledge and intuitive evaluations. In this context it becomes, however, very important to develop a statistical approach to guide the choice. The second aspect concerns the issue of quality and its assessment for multi-source production processes (Section 4). In particular, two quality aspects are briefly discussed: the evaluation of the accuracy of register estimates within ISR (Sub-section 4.1) and the strategies to improve the coherence of the estimates (Sub-section 4.2). Finally, the main lines of research for the next future are mentioned (Section 5).

In this paper most of the examples concern social variables which represent one of the most advanced examples of variables handled in a multi-source process. In this context the main sources are the Register of Individuals, the Population Census and the Social Surveys; these were re-designed more recently, for this reason the integration and the use of multi-source methods is more advanced. However, the same approach could be extended also to other kind of statistical processes.

2. The new Istat data production process

The Istat's strategic programmes focus is to integrate administrative data, create statistical registers and conduct supporting statistical surveys, in line with the new organisational, technological and methodological data production model aimed at fully exploiting all type of available data. These sets registers at the centre of the statistical data production system. The functional scheme of the new Istat production process is based on three fundamental components: ISR, PCS and surveys. The ISR constitutes the main methodological, technological and information infrastructure of the system. This component, unlike the past decade, in addition to playing a fundamental role supporting the other two components of the process, PCS and surveys, is itself an important and exclusive source for the dissemination of official statistics. The multi-source production processes applied by Istat are described in Figures 2.1 and 2.2.

The upper part of Figure 2.1 illustrates the ISR production process, based on processes of massive integration at single record level of multi-source data - mainly from administrative sources but also from surveys, as regards some missing subpopulations in the administrative sources - through the application of record linkage techniques, editing and imputation of missing data. Data production is completed through model estimation processes: micro prediction for unobserved sub-populations and correction for over and under-coverage of the registers (Pfefferman 2015). The ISR includes a metadata system that certifies the level of quality both on an overall level and in relation to its components and variables. The monitoring over time of the coverage and accuracy indicators represent the evolution of the quality level of the information made available by the ISR.

Figure 2.1 – Functional scheme of the ISR/PCS/Surveys production processes

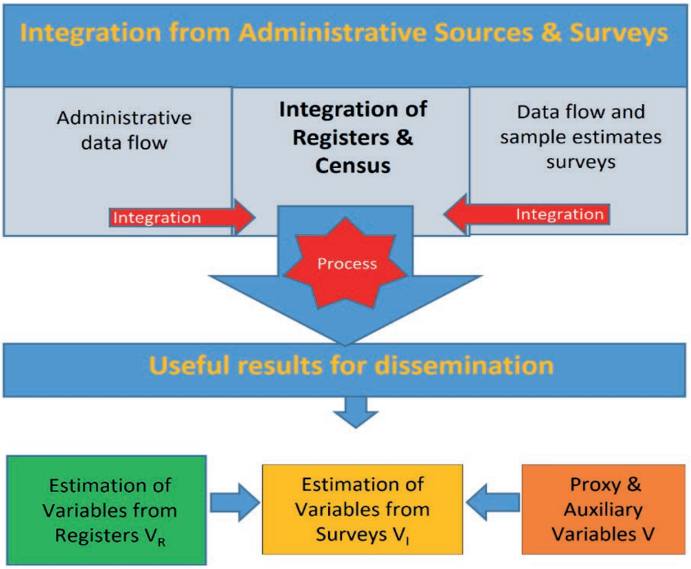
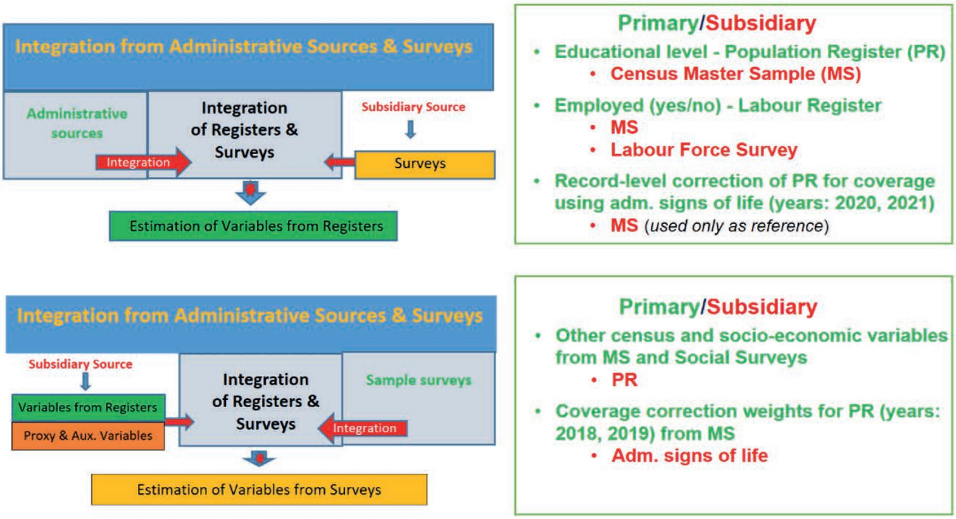

Integration from Administrative Sources & Surveys
Administrative data flow
Integration of Registers & Census
Data flow and sample estimates surveys
Integration
Integration
Process
Useful results for dissemination
Estimation of Variables from Registers V_R
Estimation of Variables from Surveys V_I
Proxy & Auxiliary Variables V

Figure 2.2 – Functional diagram of estimation in complex multi-source processes


Integration from Administrative Sources & Surveys
Administrative sources
Integration of Registers & Surveys
Subsidiary Source
Integration
Surveys
Estimation of Variables from Registers
Integration from Administrative Sources & Surveys
Subsidiary Source
Variables from Registers
Proxy & Aux. Variables
Integration of Registers & Surveys
Sample surveys
Integration
Estimation of Variables from Surveys

Primary/Subsidiary

- Educational level - Population Register (PR)
  - Census Master Sample (MS)
- Employed (yes/no) - Labour Register
  - MS
  - Labour Force Survey
- Record-level correction of PR for coverage using adm. signs of life (years: 2020, 2021)
  - MS *(used only as reference)*

Primary/Subsidiary

- Other census and socio-economic variables from MS and Social Surveys
  - PR
- Coverage correction weights for PR (years: 2018, 2019) from MS
  - Adm. signs of life

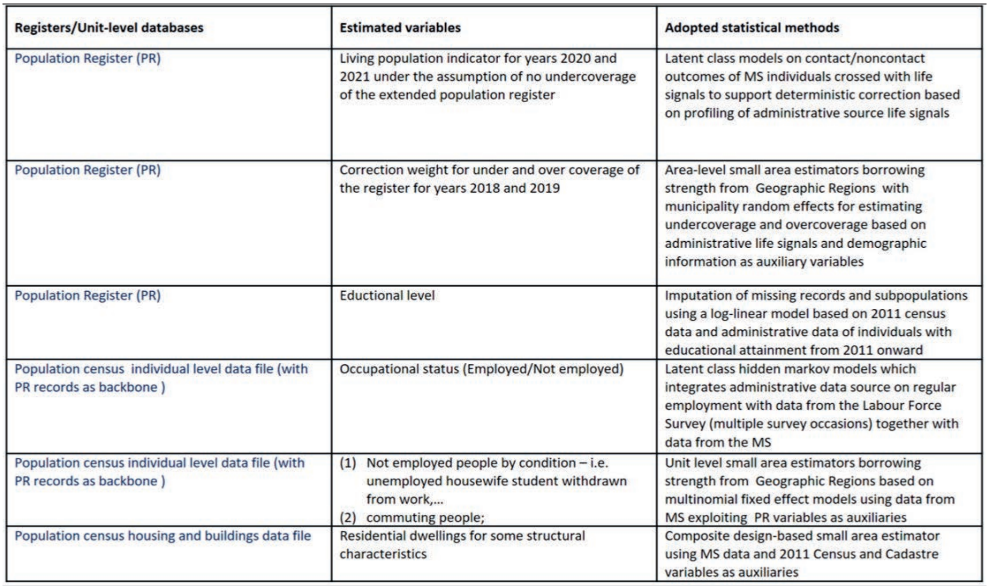
The lower part of Figure 2.1 illustrates the output production process in which the data integration into the ISR is completed and extended through the so-called direct surveys, which include in addition to the permanent censuses also the sample surveys properly systemised. For the social surveys, a structured process of bringing them into system with each other, together with the Population Census and the ISR, has been initiated. To the estimation of the variables mainly derived by administrative source, called for this reason in the following *register variables*, is added the estimation from the sample surveys of variables that cannot be deduced from administrative sources, henceforth referred to as *survey variables*. This phase may require the study and application of micro prediction techniques, called *projection estimators* (Kim and Rao 2012), which produce a vector of predicted values of the target variables for each unit belonging to the register. The result is a database containing for each unit of the register a set of predicted values that are combined with the micro information of the register for the production of the statistics of interest at the various planned and unplanned domains of interest.

Figure 2.2 focusses on the output production process in which the data integration into the ISR is completed and extended through the direct surveys; estimation becomes a complex and articulated process in which information from administrative and survey sources enter, as appropriate, on the basis of different power relations.

The upper part of Figure 2.2 illustrates the complexity of the data production process in the case of register variables. In such a situation, the main source is represented by administrative information which are good quality proxies for the target statistical variables aligned with dissemination objectives in terms of definition, classification and time reference. The observations collected through the survey, traditionally more up-to-date than the corresponding administrative information, represent a fundamental *subsidiary* source that is exploited, through the application of models, for the completion of the target variable at micro level.

The lower part of Figure 2.2 describes the data production process in the case of estimation of survey variables. This is a more traditional situation in which the survey is the main source, while data from register, together with additional information from administrative sources not included in the register (mainly proxy variables), can be exploited as micro auxiliary information (possibly information at the aggregate domain level can also be exploited in the estimation process) to improve the level of quality of the estimates produced. Again, the estimation process aims to produce micro data. Depending on the quality level of the observed data, different estimators can be studied, developed and applied. In the case where all the domains of interest are observed in the sample and the estimates are characterised by acceptable levels of quality, direct, projection model-assisted estimators are applied, which belong to the family of well-known calibration estimators that have been used for more than 30 years by Statistical Institutes. In the case where some of the domains of interest are not observed in the sample and/or the quality of direct estimates is not acceptable, indirect, projection model-based estimators are applied, which fall into the family of well-known Small Area Estimation (SAE) unit-level estimators, used with gradually increasing intensity over the past 15-20 years by the most advanced Statistical Institutes for the production of official statistics.

Table 2.1 – Incomplete list of multi-source estimation processes and methodologies



Registers/Unit-level databases	Estimated variables	Adopted statistical methods
Population Register (PR)	Living population indicator for years 2020 and 2021 under the assumption of no undercoverage of the extended population register	Latent class models on contact/noncontact outcomes of MS individuals crossed with life signals to support deterministic correction based on profiling of administrative source life signals
Population Register (PR)	Correction weight for under and over coverage of the register for years 2018 and 2019	Area-level small area estimators borrowing strength from Geographic Regions with municipality random effects for estimating undercoverage and overcoverage based on administrative life signals and demographic information as auxiliary variables
Population Register (PR)	Eductional level	Imputation of missing records and subpopulations using a log-linear model based on 2011 census data and administrative data of individuals with educational attainment from 2011 onward
Population census individual level data file (with PR records as backbone)	Occupational status (Employed/Not employed)	Latent class hidden markov models which integrates administrative data source on regular employment with data from the Labour Force Survey (multiple survey occasions) together with data from the MS
Population census individual level data file (with PR records as backbone)	(1) Not employed people by condition – i.e. unemployed housewife student withdrawn from work,... (2) commuting people;	Unit level small area estimators borrowing strength from Geographic Regions based on multinomial fixed effect models using data from MS exploiting PR variables as auxiliaries
Population census housing and buildings data file	Residential dwellings for some structural characteristics	Composite design-based small area estimator using MS data and 2011 Census and Cadastre variables as auxiliaries

Table 2.1 gives an overview of the statistical techniques applied to support the new data production process. The diagram is a not exhaustive list of the models applied to produce the different registers and other micro-level databases.

3. Choice of primary and secondary sources

In the ISR context, we need to assess whether values of a variable in a register can be directly used for computing statistics, or it is preferable using a random sample.

Roughly speaking, register data are used when information is considered of high quality and the coverage is high. These are intuitive and pleonastic considerations, and it would be useful to associate some more statistical considerations, that allow to clarify and, whenever possible, to quantify the convenience of an approach mainly based on registers.

In this context, the problem is concerned with the bias of estimates rather than their variability. There are two main issues to take into account: how much the variable in the register represents the target variable, and to which extent the observed part of the data is representative of the population. The latter problem is related to the coverage of administrative data and more in general can be set in the context of non-probability data sources used to build variables in the register. Some useful ideas on the latter issue can be drawn from Meng (2018), where there is an interesting discussion along the line of the following question "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?". More in general, this is one of the questions of a survey manager to a statistician when asking for the use of administrative data. The motivating case in the paper is referred to the LED project using unemployment insurance wage records, which cover more than 90% of the US workforce, but they exclude all federal employers. In order to use them, it is important to know how much they can help or whether they can actually do more harm than help.

To answer to the question, we need to compare the accuracy of estimates obtained with probabilistic sampling data, that are presumed to be representative of the population (property that can be violated by the non-response), and non-probabilistic sampling data that could be affected by a selection bias and in the registers are generally characterised by a number of units covering a large part of the target population. The case of an estimate of the mean value of a finite population in absence of a measurement error is discussed. Let the population of size N composed of individuals indexed with j=1,…,N. Let Y the target variable associated to the units, and let

$\bar{G}_N = \frac{1}{N}\sum_{j=1}^{N} G_j$, where $\{G_j = G(Y_j), j = 1, \dots, N\}$ the target parameter to estimate. I_n is the observed subset of units of size n from the population. The mean computed on I_n is

$$\bar{G}_n = \frac{1}{n}\sum_{j \in I_n} G_j = \frac{\sum_{j=1}^{N} R_j G_j}{\sum_{j=1}^{N} R_j}$$

where $R_j = 1$ for $j \in I_n$ and 0 otherwise. R represents the selection/observation mechanism. For probability samples, $\boldsymbol{R} = \{R_1, \dots, R_N\}$ follows a known probability distribution, differently for non-probability samples. The mechanism ruling R is determinant for the accuracy evaluation of the $\bar{G}_n$ estimator.

Starting from the equality $\bar{G}_n - \bar{G}_N = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G$, accuracy can be measured by

$$MSE_{\boldsymbol{R}}(\bar{G}_n) = E_{\boldsymbol{R}}\left[\rho_{R,G}^2\right] \times \left(\frac{1-f}{f}\right) \times \sigma_G^2 = D_I \times D_O \times D_U,$$

D_I is named as *Data Defect Index*, D_O as *Dropout odds*, D_U as *Degree of uncertainty*.

The error depends by three elements: 'data quantity' $\left(\frac{1-f}{f}\right)$, 'problem difficulty' σ_G^2, and 'data quality' that is caught from the correlation $\rho_{R,G}$. Unfortunately, although D_I is constrained by the values of D_O and D_U, it cannot be estimated from the sample, because we have only data corresponding to R=1. Nevertheless, this formula is useful for some considerations that make better understand the problem. The first question is how much we gain or lose with respect to a simple random sample taken as benchmark. We can compute the *Deff*

$$Deff = \frac{E_{\boldsymbol{R}}[\bar{G}_n - \bar{G}_N]^2}{V_{SRS}(\bar{G}_n)} = (N-1)D_I$$

This result shows that to ensure the MSE of the sample mean can enjoy of the usual converge rate of the n^{-1} order, D_I should be controlled with a rate of N^{-1}, or equivalently the 'data defect correlation' $\rho_{R,G}$ with a rate of $N^{-1/2}$. For a population of 60,000,000 of individuals (a size comparable to the Italian residents), the data defect correlation should be $\rho_{R,G} \approx 0.0001$, that is a very low value.

On the other hand, we may inspect the damage of $\rho_{R,G}$ even in the case of a small value. We may compute the effective sample size n_{eff} of the subset of non-probabilistic data to have the same MSE of a SRS:

$$n_{eff} \leq n^*_{eff} = \frac{f}{1-f} \times \frac{1}{D_I} = \frac{n}{1-f} \times \frac{1}{ND_I}.$$

We observe that ND_I increases rapidly with N causing an important reduction of n_{eff}.

For instance, with $E_R[\rho_{R,G}] = 0.05$, $n_{eff} \leq 400\frac{f}{1-f}$, then for a subset of data with f=1/2 (50% of the population) to have a behaviour similar to a SRS, the effective sample size cannot be greater than 400 units, hence with a population of 60 million units, there is a reduction of the sample size from 30 million to an effective sample size of 400 units, that is a decrease of 99.9999%, or equivalently a great loss of efficiency. The fact is that, even if D_I for a non-probabilistic sample is small, the effect is magnified by N.

We notice that if we do not control the R-mechanism, the fact that n (units observed in the register) is big can be even counterproductive, because we put more trust on biased data, as the Author says: "*The bigger the data, the surer we fool ourselves*".

So far, we have introduced some concepts useful for the choice of non-probability versus probability sample. Nevertheless, in the reality, another issue should be considered. This is the presence of non-response in the probabilistic sample survey, as stated in the original question.

With a certain approximation, an indicator considering the non-response is

$$|\rho^{BIG}_{R,G}| \leq \sqrt{\frac{D^S_O}{D^{BIG}_O}}|\rho^S_{R,G}|$$

where $D^{BIG}_O = \frac{1-f}{f}$, and $D^S_O = \frac{1-rf_s}{rf_s}$, with f and f_s denoting the sampling rate for the non-probabilistic and probabilistic samples respectively, and r is the response rate of the sample survey. One more time, this quantity cannot be directly estimated, but it can be useful to make sensitivity analysis for the case under investigation.

All the previous considerations are concerned with the bias eventually introduced by the non-representativeness of the administrative data, and are developed under the assumption that data is not affected by ‘measurement errors’, that is variables in random and non-random samples gather information without errors on the target variable. In practice, it is generally assumed that sample surveys are almost free of measurement errors while administrative data may be affected, and consequently data in the registers are compared with surveys data. This is essentially because register data are gathered for purposes different from the ones of NSIs, while surveys are designed with a specific statistical objective. This means that survey data are generally taken as a gold standard, and hence analysis of bias due to errors in register data can be carried out.

4. Quality assessment in a multi-source framework

The topic of this section concerns quality assessment of the estimates in a multi-source framework. This is a very important subject and there are many aspects that should be considered. Here we briefly introduce just two of them: uncertainty and coherence of estimates and data.

4.1 Assessing uncertainty of register-based statistics produced by multi-source data

Assessing uncertainty of statistics produced by methods integrating multi-source data of different nature is particularly relevant and is even more important when the output of inference is a statistical register, *i.e.* a set of data estimated at unit level (Zhang 2011, Zhang 2012). With statistical registers, the possibilities for making different analysis increase, so necessary information and agile tools to assess their quality must be provided to the users.

The issue of accuracy assessment with multi-source data still requires investments in terms of methodology. While in sample surveys the classical randomisation or design-based approach is generally adopted, *i.e.* the evaluation of statistical estimates is made with respect to the probability distribution induced by the sampling plan, in case of multi-source inference accuracy evaluation needs to be developed in order to deal with non-probabilistic and probabilistic data sources, often jointly used. Moreover, the need to account in the inferential stage for possible errors in data (measurement, non-observation, coverage, linkage) induces the use of statistical models.

In such a context, a first question is concerned with the sources of uncertainty to take into account when assessing accuracy. For instance, when adopting mean squared error (MSE) as a measure of accuracy, the question arises as to which random mechanisms should be taken into account in the MSE. Should it be calculated referring only to the model, to the sample, or both? Other questions are concerned with the composition of errors, is the variability of estimates still the aspect we need to focus on, or the bias that is certainly more difficult to measure? Moreover, in the context of statistical register construction where the output is a micro dataset that should represent a synthetic 'picture' of the population of interest, we wonder whether is sufficient to assess the accuracy of linear population parameters - as often done in the literature - or it is necessary to calculate the accuracy of parameters that relate to the probability distribution of the variable of interest as for instance the cumulative density function (CDF). Finally, an aspect to consider in the case of estimates from register is that the tool for assessing uncertainty should be agile, given that it should be able to evaluate unplanned estimates computed by users on the set of data available in the statistical register.

Some of these methodological issues have been recently discussed in literature. Regarding accuracy, Alleva *et al.* (2021) propose to consider model and sampling errors jointly, and address the issue of providing statistical information about the quality reported for each unit such that an appropriate combination of it allows the evaluation of accuracy of linear estimators. A study of resampling techniques for accuracy evaluation in a similar context is in Scholtus *et al.* 2021. Boeschoten *et al.* 2021 propose a multiple imputation approach applied to a data integration model-based on hidden Markov models. Other papers related to this topic can be found among the works on mass imputation, see for instance (Chen *et al.* 2022, Kim *et al.* 2021, Yang *et al.* 2021, Di Zio *et al.* 2022). Despite these studies, many aspects still need to be investigated in depth, and further methodological studies should be devoted to this end.

4.2 The integration of the Population Census and Social Surveys: Coherence by design

As described in Section 2, in recent years the information produced by Istat on the population and households has been considerably enriched in particular since 2018, when the Census began to disseminate yearly estimates about population counts and the main socio-demographic and socio-economic variables, which complement the estimates traditionally produced by social surveys. On one hand, Census estimates respond to the need for information on very detailed territorial domains, which cannot be produced by Social Surveys through direct estimates. On the other hand, the coherence issue between different figures has become increasingly relevant.

In particular, the coherence was investigated with reference to the estimates of the employment produced by Istat, as the output of different statistical production processes, each of which has specific objectives and European regulations: Labour Force Survey (LFS), Population Census (PC) and also the estimates produced by the National Accounts (NA). The same problem can also arise with reference to other variables, beyond employment.

The estimates produced by the three statistical production processes (LFS, PC, NA) have different characteristics in terms of detail and timeliness. The reference population and period also have specificities.

In particular, the LFS employment estimates (consistent with the International Labour Organization, ILO, definition, inquired through very detailed questions) are obtained through direct estimates (calibration) based on survey data. The estimates are very timely (provisional monthly estimates 30 days after the end of the reference month - national level, quarterly estimates after about 70 days - regional level, annual estimates for the year t released in March t+1 - provincial level). Administrative information on work is not used (the European regulation does not allow the imputation of the employment status at the individual level, but even if such auxiliary information could be used in the estimation phase, for example in the non-response correction, at the moment it is not used).

The PC estimates (consistent with the ILO definition as well) are obtained through an indirect estimation model. In particular, a latent class model was used, which exploits the three available sources: data collected from the PC sample surveys, LFS data, work signals from administrative sources taken from the ISR and summarised at the individual level; demographic variables, educational level, school or university attendance, fiscal administrative data (on income from work, pensions), monetary allowances are also exploited as covariates in the model. Hence, a multi-source estimation model is used to produce Census estimates. The estimate of employed persons for year t is released in December *t*+1 (municipal level).

As already mentioned, the NA employment estimates (according to the ILO definition as well) complement the picture; they are obtained integrating LFS sample with administrative data on employment, supported by probabilistic models; therefore, NA uses a multi-source estimation method as well, the main difference with the PC methodology is that administrative data are taken only for the units in the LFS sample. The estimate of employed people for the year t is released in September t+2 (provincial level).

Even taking into account the differences explained by the different reference populations and reference periods, substantial unexplained differences remain between the estimates, caused by the specificity of the statistical processes, the lack of harmonisation of the sources and the use of different estimation methodologies. The study that was conducted involved the analysis of the three employment estimates, taking into account the characteristics of the respective statistical processes, in order to identify the determinants that underlie the differences in the estimates, identifying the factors that can generate distorting effects. This study highlighted the following key aspects:

- Need for alignment of the Reference Population Frame to be used for the sample selection for PC and LFS;
- Adequacy of the PC and LFS sample designs to correctly represent the reference population. However, more efficient sample designs can be evaluated, exploiting the available auxiliary information; in this context, an integrated design could be evaluated, in which LFS constitutes a second phase survey of the Master Sample of the Population Census (this means that LFS sample would be selected as a sub-sample of the Master Sample, that is the sample of households selected from the Population Register for the Census survey);
- Bias effect due to total non-response and replacement of non-responding households in the LFS, especially in the post-pandemic period;
- Partial inadequacy of the current LFS estimation methodology to correct this bias;
- Need of studying possible improvements in the LFS estimation process, exploiting auxiliary information from the ISR in order to correct the

bias due to non-response and substitutions (or exploiting information from the PC Master Sample in the hypothesis of an integrated two-phase design);

- Need to deepen the evaluation of the measurement error of the variables collected by PC and LFS; to this aim an integrated sample design would be very useful, giving the opportunity to observe, for a subsample of individuals, the two observations in the two survey;
- Need to harmonised the auxiliary information derived from the ISR to be used to support the various estimation processes, eliminating unjustified differences and duplication in the processes (in particular, this issue concerns the differences in the ways in which the administrative sources are currently used by PC and NA);
- Need to review the statistical processes, in particular the estimation methodologies, in order to remove unnecessary differences and trying to exploit the experience already gained; this issue mainly concerns the NA estimation methodology, which must be reviewed considering the availability of statistical registers and in order to adapt the methodologies to the available sources and to the know-how gained in the field of census estimation and therefore aligning the NA estimation process with that of the Census.

In this context, the Integrated Census and Social Surveys System - ICSSS represents the prerequisite for the development and improvement of multi-source estimation methodologies, which fully exploit the potential of the available information, in order to guarantee a better quality of the estimates disseminated, also in terms of consistency as well as accuracy.

The innovations which are currently under study and which could be introduced in the LFS and PC statistical production processes, bringing the ICSSS to completion, are reported below.

As mentioned, for the LFS it is necessary to develop methodologies for correcting the bias of the estimates due to non-responses and replacement of non-respondent households. If an integrated two-phase design were used, certain variables collected by the Master Sample on the overlapping sample could be exploited in the LFS estimation process, in the context of methodologies for the correction of total non-response. In particular, methodologies based on simple post-stratification or calibration of the sample, according to totals estimated by the PC can be adopted. In the case of calibration, the perspective would be that of rotated samples, in which the Master Sample represents the first wave of data collection, previous to LFS. In this context, the following items must be duly considered: the portion of the sample that was not selected from the Master Sample, the discrepancies between the non-responses to the PC and LFS and the different reference population compared to the census population.

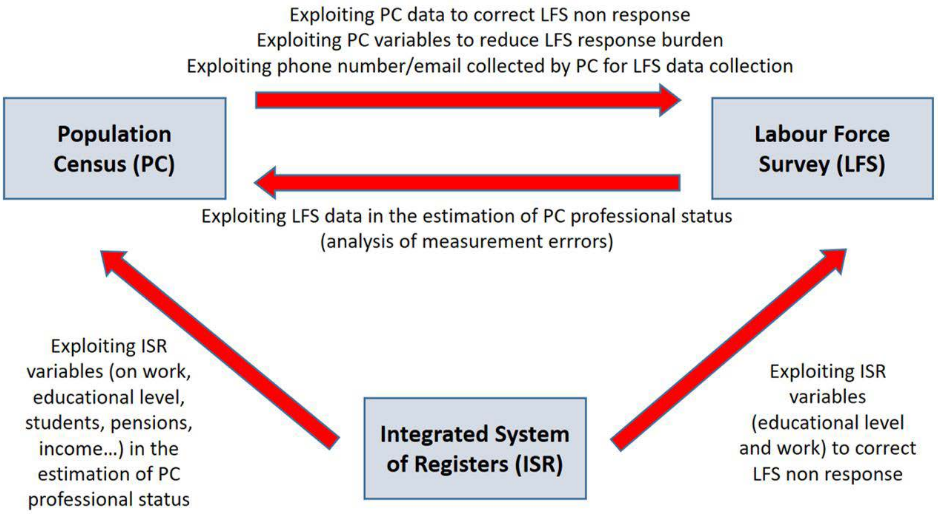
More generally, in the LFS estimation process, the exploitation of variables derived by the ISR is currently being studied (in addition to the demographic variables already used in the calibration). In this context, the employment status from the Labour Register and the educational level from the Register of the Individuals are relevant variables to be introduced as auxiliary information in the estimation, as they are strongly correlated with the target variable. The preliminary results show that the use of both of these administrative variables actually makes it possible to almost completely correct the bias generated by non-responses and replacements of non-responding households and improves the efficiency of the estimates; at the same time, the exploitation of administrative variables, which represent an important source of information in the PC estimate, makes it possible to achieve better consistency between LFS and PC estimates.

From the PC point of view, as already mentioned, in the models to estimate the employment condition (latent class models) microdata from the LFS already represent one of the three available measures of the latent variable, in addition to PC and administrative data. In particular, at present it has been assumed that LFS collected data are perfect measurements and are not affected by measurement errors.

There is currently no overlap between LFS and PC. In perspective, the integration of LFS as a second phase survey of the Master Sample would guarantee an overlap of a portion of the PC sample with a portion of the LFS sample and therefore, for these observations, the three measures will be available. The availability of the three measures would make it possible to study the measurement error of the surveys with the aim of taking it into account in the estimation process, improving the coherence.

Figure 4.1 synthetically shows the integration of the PC and LFS statistical processes, in the context of an ICSSS, exploiting also the information from the ISR; here, the reference target variable is the employment status, for which the goal of a better coherence between the different estimates disseminated by Istat is a very relevant issue. However, a similar process could also be extended to other variables of interest, measured by the PC and a generic Social Survey.

Figure 4.1 - The integration of PC and LFS statistical processes


Exploiting PC data to correct LFS non response
Exploiting PC variables to reduce LFS response burden
Exploiting phone number/email collected by PC for LFS data collection
Population Census (PC)
Labour Force Survey (LFS)
Exploiting LFS data in the estimation of PC professional status (analysis of measurement errrors)
Exploiting ISR variables (on work, educational level, students, pensions, income...) in the estimation of PC professional status
Integrated System of Registers (ISR)
Exploiting ISR variables (educational level and work) to correct LFS non response

5. Main lines of research for the next few years

To conclude, in this last section we introduce the main lines of research in Istat concerning the integrated multi-source statistical production processes, so far opened in Istat.

Starting with the new data production process, the main activities and challenges for the near future are the following:

- Completing the methodological framework and implementation of the registers in the ISR;
- Defining the final asset of the second component of the CSSIS regarding Social Surveys and their coordination with the Census Master Sample. In particular, it will be necessary to define which Social Surveys will be second phase surveys from the Master Sample.

Finally, we highlight some relevant general methodological issues in an integrated process design, already mentioned in previous pages:

- Design and implementation of quality by design in estimation processes with special reference to the efficiency of the estimates produced;
- Consistency between estimates produced by different processes and different sources (such as direct surveys, administrative sources but also from big data sources);
- Timeliness of estimates in terms of minimum distance from the reference period;
- Level of disaggregation of statistics produced at the level of spatial and structural detail;
- Joint exploitation, even over time, of data from different sources to produce new information with greater timeliness and/or strong granularity.

References

Alleva, G., P.D. Falorsi, F. Petrarca, and P. Righi. 2021. "Measuring the Accuracy of Aggregates Computed from a Statistical Register". *Journal of Official Statistics - JOS*, Volume 37, Issue 2: 481-503.

Chen, S., S. Yang, and J.K. Kim. 2022. "Nonparametric Mass Imputation for Data Integration". *Journal of Survey Statistics and Methodology*, Volume 10, Issue 1: 1-24.

Di Zio, M., R. Filippini, and S. Toti. 2022. "Variance estimation for the mass imputation of the "Attained level of education" in the Italian Base Register of individuals: A comparison between analytical and MonteCarlo estimates". Paper presented at the *Expert Meeting on Statistical Data Editing*. 3rd-7th October 2022 (virtual), United Nations Economic Commission For Europe Conference of European Statisticians. https://unece.org/statistics/events/SDE2022.

Kim, J.K., S. Park, Y. Chen, and C. Wu. 2021. "Combining Non-Probability and Probability Survey Samples Through Mass Imputation". *Journal of the Royal Statistical Society, Series A: Statistics in Society*, Volume 184, Issue 3: 941-963.

Kim, J.K., and J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach". *Biometrika*, Volume 99, Issue 1: 85-100.

Meng, X-L. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Elections". *The Annals of Applied Statistics*, Volume 12, Issue 2: 685-726.

Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture". *Journal of Survey Statistics and Methodology*, Volume 3, Issue 4: 425-483.

Scholtus, S., and J. Daalmans. 2021. "Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data". *Journal of Official Statistics - JOS*, Volume 37, Issue 2: 433-459.

Yang, S., J.K. Kim, and Y. Hwang. 2021. "Integration of data from probability surveys and big found data for finite population inference using mass imputation". *Survey Methodology*, Volume 47, Issue 1: 29-58.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*, Volume 66, Issue 1: 41-63.

Zhang, L.-C. 2011. "A Unit-Error Theory for Register-Based Household Statistics". *Journal of Official Statistics - JOS*, Volume 27, Issue 3: 415-432.

A study of MLP for the imputation of the "Attained Level of Education" in Base Register of Individuals

Fabrizio De Fausti[1], *Marco Di Zio*[1], *Romina Filippini*[1],
Simona Toti[1], *Diego Zardetto*[2]

Abstract

The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of administrative information. Nevertheless, it is needed to resort to sample survey data to cope with delay of information and coverage problems. Due to the complexity and heterogeneity of the available information, the solution of the problem with standard statistical methods needs the construction of different imputation models with a strong effort in terms of human intervention. We study the use of a multilayer perceptron model to make the process more automatic, i.e. less costly in terms of human resources, and possibly more accurate in terms of estimates. Since a relevant quality aspect is the ability of the imputation process to provide a good estimate of the ALE frequency distribution, sampling weights are used in both classic statistical techniques and in the context of machine learning.

Keywords: Register-based statistics, imputation, data integration, sampling weights, machine learning.

1. Introduction

The Italian production system of statistics is deeply changing, moving towards a register -based statistics production. The Base Register of Individuals (BRI), resulting from the integration of data from different sources, is the basis of the Permanent Italian Census that is as much as possible register-based. Among others, the Italian Census gathers information on the Attained Level of Education (ALE). The high amount of available information on this topic,

[1] Fabrizio De Fausti (defausti@istat.it); Marco Di Zio, (dizio@istat.it); Romina Filippini (filippini@istat.it); Simona Toti (toti@istat.it), Italian National Institute of Statistics - Istat.

[2] Diego Zardetto (dzardetto@worldbank.org, and zardetto@istat.it), World Bank.

in particular longitudinal information, may allow the production of statistics from register rather than from survey. The aim is to insert the variable ALE in the set of BRI core information. In particular, we are interested in a micro level estimation of the ALE (8 classes) for Italian resident population.

However, delay of information and coverage problems arise, hence micro data imputation/prediction is necessary. Due to the complexity and heterogeneity of the available information, the solution of the problem with standard statistical methods requires an in-depth knowledge of data structure and an expensive initial phase of data analysis and treatment. Moreover, different imputation procedures must be combined to deal with sub-populations characterised by different amounts of information (Di Zio *et al.*, 2019; Di Zio *et al.* 2018).

In the last years, machine learning (ML) techniques have been applied in many contexts (including official statistics, see Yung *et al.* 2018) with the aim of improving predictions, especially when very large collections of data can be leveraged. The advantage of using such techniques is also in their almost automated application to data. These opportunities motivated the study of ML for the ALE prediction task, with the twofold objective of improving estimation accuracy (given the high amount of available data) and reducing human workload (given the efforts needed for data treatment and sequential usage of complex models in the official procedure).

In recent years, Istat has gained considerable experience in the use of neural networks to extract statistical information from extremely large and unstructured datasets generated by non-traditional sources. In particular, the potentialities of deep-learning models like convolutional neural networks (CNN) have been investigated (Bernasconi *et al.* 2022, De Fausti *et al.*, 2020, De Fausti *et al.*, 2020) for the treatment of images and natural language. In this study, we focus instead on the Multilayer Perceptron model (MLP).

Early applications of the MLP model in the field of official statistics can be found in (Nordbotten *et al.* 1995, Nordbotten *et al.* 1996, Charlton *et al.* 2004).

In a point of view of modernisation of statistics there is a strong push in introducing ML algorithms into the production processes; the High-Level Group for the Modernisation of Official Statistics of UNECE (HLG-MOS) launched a Machine Learning project[3] in 2019 with the aim of facilitating the investigation of the use of machine learning for official statistics. The project aimed to investigate whether machine learning can help produce more relevant, timely, accurate and reliable estimates. Studies of this type have revealed the need for a broader methodological framework that concerns important dimensions of the quality of ML algorithms such as Accuracy, Explicability, Reproducibility, Timeliness, Cost-effectiveness. The fields of applicability that have been investigated are Coding and Classification, Edit and Imputation and Imagery Analysis. For these fields and in general it emerged that the ML algorithms "are a must where they can add value, and they should not be used where it does not" (United Nations Economic Commission for Europe 2022). Especially for classification and coding, studies have shown that machine learning can provide better quality results than a strictly manual method in terms of timeliness (and indirectly in terms of cost).

As part of the work of the UNECE group on the effectiveness of ML as a tool for improving and modernising imputation processes, Istat worked on a comparison between the official imputation approach for ALE estimation, based on log-linear models, and the Multilayer perceptron models (De Fausti *et al.*, 2022).

The evaluation focusses on two quality aspects: accuracy of predictions (and of estimated aggregates computed by directly using the predictions) and efficiency of the procedure. The efficiency assessment is primarily concerned with the automation of the process, which means that resources spent for data analysis and preparation can be minimised. Results are encouraging especially concerning the efficiency. In fact, we do not notice an improvement in terms of accuracy, but the same level of quality is reached by using raw data, that is without resorting to expensive data pre-treatment steps. In that application, survey data are used without considering sampling weights.

An important quality dimension of the imputation process is the accuracy of the imputed variable ALE at micro level. Another relevant quality aspect is the ability of the imputation process to provide a good estimate of the ALE frequency distribution. The latter is a crucial goal for Istat since the ALE distribution on the Italian resident population will be a standard output of the

[3] https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project.

new yearly Permanent Italian Census. In this work, we extend the study in (De Fausti *et al.*, 2022) to include sampling weights in Multilayer perceptron models. The role of sampling weights is to make the sample representative of the whole population, thereby leading to unbiased estimates. Although still under discussion, techniques to incorporate sampling weights in classical statistical models are developed (Pfefferman, 1993), the same cannot be said for machine learning models.

The paper is structured as follows. In Section 2, we briefly explain how sampling weights are taken into account in a survey and in a ML approach; Section 3 describes the data used for our experimentation; Section 4 describes the official and the machine learning imputation methods compared in this study using the sampling weights; Section 5 describes the experimental study; some final remarks are given in Section 6.

2. Taking into account sampling weights

2.1 Sampling weights in surveys

National Statistical Institutes (NSIs) routinely use complex sampling designs to carry out probability sample surveys. This practice results from the need to find a tradeoff between statistical efficiency and logistic constraints. To the scope of the present paper, any sampling design resulting in unequal inclusion probabilities of the observed sample units can be considered complex. Any statistical analysis on complex survey data should be performed taking into account the selection of sample units with unequal probabilities. Failing this, inferential results would be generally invalid, even under ideal conditions (that is neglecting any form of non-sampling error, like sampling frame imperfections, non-response, measurement errors, etc.). The design-based/model-assisted approach to finite population sampling is the reference inferential framework adopted by NSIs. By properly incorporating inclusion probabilities into estimators, it leads to unbiased (or asymptotically unbiased in the large sample limit) estimation without any need of model assumptions on the target population. In this approach to inference, inclusion probabilities typically enter estimator expressions in the form of weights attached to survey units. Horvitz-Thompson estimators use so-called design weights, which are reciprocals of inclusion probabilities. Calibration estimators, which leverage available auxiliary information on the target population to improve estimation efficiency, employ so-called calibration weights, which are complex non-linear functions of design weights and embedded auxiliary information. Furthermore, despite non-sampling errors mark a departure from the ideal conditions underpinning the validity of the design-based/model-assisted inferential framework, NSIs invariably strive to mitigate estimation flaws that could arise from non-sampling errors by adjusting the weights. For instance, to adjust weights for total non-response and/or frame imperfections, propensity score modelling and calibration are commonly applied alternatives, the choice among the two being mainly driven by the available auxiliary information.

Put briefly, the joint effect of (i) unequal inclusion probabilities and (ii) usage of auxiliary information for survey estimation determines unequal survey weights that should not be overlooked when fitting statistical models to data from complex surveys.

2.2 Sampling weights in ML

To the best of our knowledge, the question of whether (and how) survey weights should be incorporated in Machine Learning models trained to survey data has received little to null attention in the literature. One possible explanation might be that research in the ML field is typically more concerned with achieving high prediction accuracy at micro-level than aimed at obtaining reliable estimates of model and/or finite population parameters. However, as explained in the introduction, the latter objective is of major relevance to the scope of our work. In fact, we need to assess the ability of MLP imputation to provide good estimates of the ALE frequency distribution, which is one of the standard statistics disseminated by the Italian Permanent Census on a yearly basis. For this reason, in order to leverage survey weights during the training phase of our MLP, we used a loss function weighted with sampling weights.

The intuition behind this formula is similar to the "census equations" leading to the Pseudo Maximum Likelihood (PML) approach. Basically, the loss function is computed on a pseudo-population of training observations obtained by cloning each training example i, wi (weight) times. This way, the MLP incurs different misclassification costs for different training examples, owing to unequal survey weights. As compared to the ordinary unweighted loss, the expected effect is to improve classification results of the MLP especially for groups of survey units characterised by higher-than-average weights. In turn, this could be particularly beneficial to MLP predictions for low frequency ALE classes, which might be under-represented in the unweighted sample

3. Data description

3.1 Resident population data

ALE for the Italian resident population in 2018 is estimated by using administrative data, traditional Census data and sample survey data.

Administrative data: administrative information on ALE is gathered making use of the information collected by the Ministry of Education, University and Research (MIUR). MIUR provides information about ALE and course attendance for people entering a study programme after 2011 and covers the period from 2011 to 2017 (scholar year 2017/2018).

Traditional Census data (2011 Census): for people that have not attended any course since 2011 we turn to data from 2011 Census to fill the gap.

Sample survey data: direct measurement for ALE in 2018 is available only for a subset of population (about 5%), coming from the first Permanent Census Survey that took place in Italy in October 2018 (CS2018).

The structure of available information is summarised in Table 3.1. Grey cells indicate that the information is not available for the specific subpopulation.

Table 3.1 - Structure of available information for mass-imputation of the attained level of education at time *t*

Source;	BRI	MIUR	2011 CENSUS	Sample		
Available inf.:	Core inf.	ALEt-1	ALEt-1	ALEt	Group	Used in the case study
Coverage			▨		A	Yes
			▨	▨	A	No
		▨			B	Yes
		▨		▨	B	No
		▨	▨		C	Yes
		▨	▨	▨	C	No

Core information like age, gender, citizenship, marital status, place of birth and place of residence are available for all individuals.

The different availability of information on ALE from 2011 to 2017, determines the partition of our population of interest into three subgroups:

A. All persons for whom information on ALE is available from MIUR belong to subgroup A;

B. Persons not in MIUR who were interviewed in the 2011 Census belong to subgroup B. This means that subgroup B is made up of individuals for whom the only information on ALE comes from the 2011 Census;

C. Individuals neither in MIUR nor in 2011 Census belong to group C. For this group no information on ALE is available.

The classification adopted for ALE is composed by 8 items: 1 – Illiterate, 2 - Literate but no formal educational attainment, 3 - Primary education, 4 - Lower secondary education, 5 - Upper secondary education, 6 - Bachelor's degree or equivalent level, 7 - Master's degree or equivalent level, 8 - PhD level.

4. Methods

4.1 Official procedure: Log-linear model imputation

The adopted official procedure is based on log-linear imputation. As stated by Singh (1988), this method generalises hot-deck imputation by choosing suitable predictors for forming "optimal" imputation classes. In fact, the approach is based on modeling the associations between variables.

The general idea is to estimate a model for the prediction of ALE at time t (I_t) - given the values of known covariates X. In particular, we estimate the conditional probabilities $h(I_t|X)$ and then impute I_t by randomly taking a value from this distribution. The conditional probabilities $h(I_t|X)$ are estimated by means of log-linear models as follows. First, a log-linear model is applied to the contingency table obtained by cross-classifying the variables (I^t, X) to estimate their expected counts $\widehat{N}(I_t|X)$ from which we can compute the counts $\widehat{N}(X)$. The estimated conditional probability distribution $\hat{h}(I_t|X)$ is easily obtained by computing $\widehat{N}(I_t|X)/\widehat{N}(X)$. This approach includes as a special case the random hot deck when a saturated log-linear model is assumed, but it has the advantage of allowing the use of more parsimonious model as well. This is an important characteristic especially when the number of variables and of the contingency table cells increases.

In order to consider sampling weights in the model, it is adopted a pseudo-maximum likelihood approach that consists in estimating log-linear models on weighted count data (Thibaudeau *et al.*, 2017, Skinner *et al.*, 2010).

4.2 Machine learning procedure: Multilayer perceptron model

We apply the MLP for the mass imputation of ALE with the same categorical input variables of the log-linear models.

Our approach aims to be as general as possible, therefore:

a. We train a single neural network, unlike the standard approach, where different models are built, according to the variables available for each profile;
b. We encode the input variables of the perceptron multilayer as one-hot encoding, in this representation the missing value of a variable is encoded like any other mode of the variable;
c. We encode the input variables of the perceptron multilayer with the aim of minimising the cross-entropy loss function. The cross-entropy is a measure of the distance between the distribution of the output variable and the distribution of the target variable.

In order to leverage survey weights during the training phase of our MLP, we modified the cross-entropy loss function as follows:

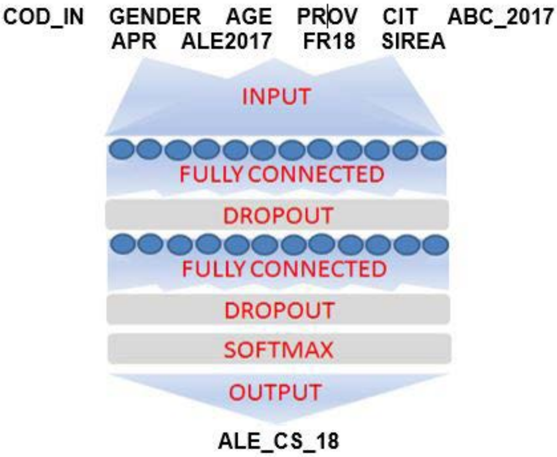
$$loss_w = -\sum_{ic} w_i T_{ic} \log(P_{ic})$$

where w_i is the final survey weight of the i-th training observation, c is the modality index in a one-hot representation, T_{ic} is the ground-truth value of the target variable for the i-th observation, and P_{ic} is the corresponding softmax function output probability distribution of the MLP.

The architecture of the network is shown in Figure 4.1 and has two hidden layers each of 128 neurons, an output layer with 8 neurons (one per modality of the target variable). To limit the over-fitting in the learning phase, two layers of dropout have been interposed. The best configuration of some hyper-parameters (number of hidden neurons, dropout probability, learning rate) was explored through a suitable grid search.

For each record of the dataset, the model generates a probability distribution on the 8 ALE items. In a conventional ML approach, the imputed value is the modal value of the distribution. However, in our case study, an important goal is to reproduce the distribution of ALE in the population of interest. To increase the distributional accuracy, for each record we impute the ALE item randomly extracted from the probability distribution of the correspondent pattern as in the log-linear models.

Figure 4.1 - Architecture of the model implemented



5. Experimental study

5.1 Description of the simulation

The comparison of MLP with the official imputation model was carried out on the 312,813 people residents in Lombardia in 2018 with no missing data on ALE 2018. The target variable is the self-declared ALE in the 2018 sample census, referring to the year 2018, which corresponds approximately to 5% of the total population of interest. The subset of units is limited to the subpopulations B and C as classified in Table 3.1.

A first experiment is carried out by using the MLP with the same covariates selected for log-linear models. The goal is to minimise confounding factors, therefore allowing for a neat comparison of results in terms of statistical accuracy.

In a second experiment (MLP all-in), data provided to MLP are not pre-processed. All the variables in the dataset enter the MLP algorithm without any selection or reclassification. In particular, the variables age and citizenship are not aggregated into classes and the variables relating to the type of school attended are used as they are presented from administrative sources, without any type of aggregation. The variables relating to the place

Table 5.1 - Variables in the dataset used in the three log-linear models and MLP approach

Id	NAME	DESCRIPTION	Log-linear			MLP	MLP without pre-processing
			A	B	C		
1	COD_IND	Record id					
2	GENDER	Gender		1	1	1	1
3	AGE_CLASS	Age classified into 14 levels	1	1	1	1	
4	AGE	Age in years					1
5	BIRTH_MU	Municipality of birth					1
6	BIRTH_CO	Country of birth					1
7	MUN	Municipality of residence					1
8	PROV	Province of residence		1		1	1
9	CIT_CLASS	Citizenship (Italian/Not Italian)	1	1	1	1	
10	CIT	Country of citizenship					1
11	ABC_2017	Subpopulation (A, B, C)				1	
12	APR	ALE from APR classified into 4 levels			1	1	1
13	ALE2017	2017 ALE (combination of Administrative and 2011 Census)	1	1		1	1
14	FR18_CLASS	Aggregated type of school and year of attendance in 2017/2018	1			1	
15	FR18	Type of school and year of attendance in 2017/2018					1
16	SIREA	Resident in Italy in 2011 not caught by the 2011 Census			1	1	1
17	ALE_CS18	2018 ALE from 2018 Census Survey				Target variable	

of residence and place of birth are also included. Moreover, the information on the data source of the three subpopulations is not considered and the flag variable (ABC_2017) which identifies the three subgroups A, B, and C, is not introduced. This second experiment is clearly meant to study the possibility of using a more automated approach for the prediction of the ALE variable in large-scale production settings.

The variables used in the different experiments are described in Table 5.1

The results of estimates obtained with MLP are compared with those of the official procedure. Quality measures are concerned with predictive accuracy of each unit and accuracy of estimated aggregates (quantities obtained by aggregating the unit predictions). The first measure is generally the one analysed in ML approaches, while the second is usually considered in National Statistical Institutes when evaluating the quality of an estimation procedure. Since the ALE distribution will be published by gender, age classes, and citizenship, it is important to evaluate the distributional accuracy in these specific subpopulations. The aggregates considered in this study refer to the main figures that are officially disseminated by Istat. In particular, we report results for the ALE distribution by citizenship.

Accuracy is calculated using a k-fold approach with k=5. The database is partitioned into 5 subgroups and:

a. the model is estimated on the training set, consisting of 4 of the 5 subgroups;
b. the results are applied on the test set, composed of the remaining subgroup;
c. accuracy is calculated only on the test set as the difference between the estimated ALE 2018 and the observed ALE 2018.

Tasks 1-3 are repeated 5 times so to reconstruct the entire dataset. The same approach is used for both ML and log-linear models so that results can be compared.

After the implementation of this approach, each individual (in each k-fold) has two probability distributions on the 8 ALE items, estimated using ML and log-linear models. The imputation process consists of extracting a random value from the probability distribution. The same imputation process is repeated 100 times to consider the model variability and the resulting indicators are averaged over those repetitions.

5.2 Results

Table 5.2 shows the micro-level predictive accuracy attained by log-linear and MLP approaches. For each method and k-fold, the proportions of units with predicted ALE equal to the observed (*i.e.* true) value are reported as percentages.

Table 5.2 - Micro-level accuracy in the 5 test sets averaged over 100 runs: Log-linear, MLP estimation and MLP All-in (percentage values)

K-fold	Log-linear	MLP	MLP All-in
1	71.202	71.521	73.047
2	71.254	71.648	73.059
3	71.155	71.350	73.209
4	71.183	71.405	73.279
5	71.023	71.385	73.155
Mean	71.163	71.462	73.150
Standard Deviation	0.077	0.110	0.088

Source: Authors' processing

The results of the MLP are very similar to those that originated from log-linear models: the average predictive accuracy, computed over the 5 folds, are respectively equal to 71.2% and 71.5%. MLP all-inn has a slightly better behaviour.

To evaluate the performance of the imputation procedures at macro level, the estimated frequency distribution of ALE in 2018 is compared with that computed using the 2018 census sample.

A possible synthetic measure is given by the Kullback-Leibler (DKL) divergence.

$$D_{KL}(T|\hat{T}) = \sum_{c=1}^{K} T_c \log_2\left(\frac{T_c}{\hat{T}_c}\right)$$

It measures the divergence of the distribution from , or, in other words, the information lost when is used to approximate T. If the two distributions are identical the Kullback-Leibler divergence is equal to 0.

Table 5.3 provides the DKL computed for log-linear, MLP and MLP all-in estimation methods. Also in macro accuracy, MLP is very close to log-linear imputation, while MLP all-in shows a greater difference.

Table 5.3 - Macro-level accuracy: Kullback-Leibler divergence (DKL) in the 5 test sets averaged over 100 runs: Log-linear, MLP estimation and MLP All-in

K-fold	Log-linear	MLP	MLP All-in
1	0.008	0.019	0.022
2	0.017	0.014	0.045
3	0.015	0.044	0.057
4	0.032	0.018	0.114
5	0.024	0.020	0.102
Mean	0.019	0.023	0.068
Standard Deviation	0.008	0.011	0.035

Source: Authors' processing

Table 5.4 reports DKL for the distribution of ALE 2018 by citizenship. We notice that the largest differences are related to the subpopulation of ‘not Italian’. This subpopulation is much smaller than the Italian one, consisting of about 27 thousand individuals (less than 9% of total population analysed), and less information is available for it.

Within the Italian subpopulation, we notice that MLP and log-linear have a similar performance to the previous tables, with a small preference for log-linear. On the other hand, it is interesting to note that MLP has a better performance in the ‘not Italians’.

Table 5.4 - Kullback-Leibler divergence between Estimated and target ALE 2018 distribution by citizenship: Log-linear *vs.* MLP *vs.* MLP All-in estimation (test set 2 averaged over 100 runs)

ALE in 2018	Italian			Not Italian		
	Log-linear	MLP	MLP All-in	Log-linear	MLP	MLP All-in
Illiterate	0.029	0.023	-0.014	0.093	0.206	-0.080
Literate but no ed. Att.	-0.014	0.025	0.047	-0.829	0.226	-0.480
Primary education	-0.176	-0.071	-0.181	0.103	-0.654	-0.262
Lower secondary ed.	0.043	-0.075	-0.757	0.479	-0.115	2.671
Upper secondary ed.	0.148	0.151	0.965	1.385	0.361	0.774
Bachelor’s degree	0.002	-0.021	-0.204	-1.249	-0.881	-1.726
Master’s degree	0.021	0.032	0.259	0.585	1.273	0.053
PhD	-0.043	-0.053	-0.079	-0.133	-0.090	-0.256
D_{KL}	0.009	0.011	0.035	0.433	0.325	0.694

Source: Authors' processing

6. Final remarks and future developments

This paper aims at investigating the behaviour of MLP as a tool for improving quality and efficiency of the statistical process of ALE estimation. A comparative study with the officially adopted log-linear imputation procedure is carried out. In order to leverage survey weights during the training phase of our MLP we modified the cross-entropy loss function using the sampling weights to create a pseudo-population. The effect is to improve classification results, especially for groups of survey units characterised by higher-than-average weights and for low-frequency ALE classes, which might be under-represented in the unweighted sample. For the imputation of ALE the results of the MLP are very similar to those originated from log-linear models in terms of predictive accuracy and macro-level estimated frequency distribution. Another important aspect that should be studied in the future is the management of longitudinal information in both MLP and standard approach in order to obtain consistent estimates over time.

This study encourages to deepening the opportunity given by the use of ML methods of a more automated approach for the prediction of the ALE variable in large-scale production settings.

References

Bernasconi, E., F. De Fausti, F. Pugliese, M. Scannapieco, and D. Zardetto. 2022. “Automatic extraction of land cover statistics from satellite imagery by deep learning”. *Statistical Journal of the IAOS*, Volume 38, Issue 1: 183-199.

Charlton, J. 2004. “Editorial: Evaluating Automatic Edit and Imputation Methods, and the EUREDIT Project”. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, Volume 167, Issue 2: 199-207.

Chen, S., D. Haziza, C. Léger, and Z. Mashreghi. 2019. “Pseudo-population bootstrap methods for imputed survey data”. *Biometrika*, Volume 106, Issue 2: 369-384.

Daalmans, J. 2017. “Mass imputation for Census estimation”. *Discussion Paper*, 2017-04. The Hague/Heerlen, The Netherlands: Statistics Netherlands - CBS.

De Fausti, F., M. Di Zio, R. Filippini, S. Toti, and D. Zardetto. 2022.

"Multilayer perceptron models for the estimation of the attained level of Education in the Italian Permanent Census". *Statistical Journal of the IAOS*, Volume 38, Issue 2: 637-646.

De Fausti, F., F. Pugliese, and D. Zardetto. 2020. "Automated Land Cover Maps from Satellite Imagery by Deep Learning". In Pollice, A., N. Salvati, and F. Schirripa Spagnolo (*Eds*.). *Book of Short Papers*. SIS 2020: 242-247. London, UK: Pearson.

De Fausti, F., F. Pugliese, and D. Zardetto. 2020. "Towards automated website classification by Deep Learning". *Rivista di statistica ufficiale/ Review of official statistics*, N. 3/2020: 9-50. Roma, Italy: Istat. https://www.istat.it/en/archivio/271059.

de Waal, T. 2016. "Obtaining numerically consistent estimates from a mix of administrative data and surveys". *Statistical Journal of the IAOS,* Volume 32, Issue 2: 231-243.

Di Cecco, D., D. Di Laurea, M. Di Zio, R. Filippini, P. Massoli, and G. Rocchetti. 2018. "Mass imputation of the attained level of education in the Italian System of Registers". Document presented at UNECE *Workshop on Statistical Data Editing*. Neuchâtel, Switzerland, 18th - 20th September 2018. https://unece.org/statistics/events/SDE2018.

Di Consiglio, L., M. Di Zio, and D. Filipponi. 2019. "An empirical evaluation of latent class models for multi-source statistics". Presentation at the 6th *Italian Conference on Survey Methodology - ITACOSM 2019*. Firenze, Italy, 5th - 7th June 2019.

Di Zio, M., R. Filippini, and G. Rocchetti. 2019. "An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data". *Rivista di statistica ufficiale/Review of official statistics*, N. 2-3/2019: 143-174. Roma, Italy: Istat. https://www.istat.it/en/archivio/271219.

Lumley, T., and A. Scott. 2017. "Fitting Regression Models to Survey Data". *Statistical Science*, Volume 32, Issue 2: 265-278.

Nordbotten, S. 1996. "Neural Network Imputation Applied to the Norwegian 1990 Population Census Data". *Journal of Official Statistics - JOS*, Volume 12, Issue 4: 385-401.

Nordbotten, S. 1995. “Editing Statistical Records by Neural Networks”. *Journal of Official Statistics - JOS*, Volume 11, Issue 4: 391-411.

Pfeffermann, D. 1993. “The Role of Sampling Weights When Modeling Survey Data”. *International Statistical Review/Revue Internationale de Statistique*, Volume 61, Issue 2: 317-337.

Runci, M.C., G. Di Bella, and F. Cuppone. 2017. “Integrated Education Microdata to Support Statistics Production”. In Lauro, N.C., E. Amaturo, M.G. Grassia, B. Aragona, and M. Marino (Eds.). *Data Science and Social Research: Epistemology, Methods, Technology and Applications*: 283-290. New York, NY, U.S.: Springer, *Studies in Classification, Data Analysis, and Knowledge Organization*.

Scholtus, S. 2018. “Variances of Census Tables after Mass Imputation”. *Discussion Paper*, December 2018. The Hague/Heerlen, The Netherlands: Statistics Netherlands - CBS.

Scholtus, S., and J. Pannekoek. 2015. “Mass-imputation of educational levels (in Dutch)”. *Internal Report*. The Hague/Heerlen, The Netherlands: Statistics Netherlands - CBS.

Singh, A.C. 1988. “Log-linear imputation”. *Working Paper, Methodology Branch*, N. SSMD-88-029 E. Ottawa, Canada: Statistics Canada.

Skinner, C.J., and L.-A. Vallet. 2010. “Fitting Log-Linear Models to Contingency Tables from Surveys with Complex Sampling Designs: An Investigation of the Clogg-Eliason Approach”. *Sociological Methods & Research*, Volume 39, Issue 1: 83-108.

Thibaudeau, Y., E. Slud, and A. Gottschalck. 2017. “Modeling Log-Linear Conditional Probabilities for Estimation in Surveys”. *The Annals of Applied Statistics*, Volume 11, Issue 2: 680-697.

United Nations Economic Commission for Europe - UNECE. 2021. *Machine Learning for Official Statistics*. Geneva, Switzerland: United Nations. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf.

Yung, W., J. Karkimaa, M. Scannapieco, G. Barcaroli, D. Zardetto, J.A.R. Sanchez, B. Braaksma, B. Buelens, and J. Burger. 2018. “The Use of Machine Learning in Official Statistics”. *Report of the UNECE Machine Learning Team*. Geneva, Switzerland: UNECE.

SESSION 3

Methodologies for big data

INTRODUCTION

Li-Chun Zhang[1]

Zhang and Haraldsen (2022, Table 1) group and exemplify the various kinds of data that fall within the scope of this session as follows, which include all the cases that will be discussed today

- *Register*: vital events, diagnoses, wage, income tax, VAT, welfare payments;
- *Transaction*: scanner data price, point-of-sales receipt, bankcard or giro pay- ment, B2B or B2P invoice, property sales contracts, ownership registration;
- *Remote sensing, with fixed sensors*: smart metre readings, weather station readings, traffic loop signals;
- *Remote sensing, with "mobile" sensors*: satellite/drone images, airborne laser scanning, maritime AIS, lorry tracking signals, mobile phone signals;
- *Internet*: web pages, social media posts.

Making use of such data for official statistics may present difficult challenges in three key aspects:

- Access (due to sensitivity, regulatory framework and public acceptance);
- Process (from input data to statistical outputs);
- Assess (of output accuracy).

Very often, there arises a need to conceptualise or revise the appropriate target parameters or statistical definitions, while wrestling with these challenges. First, the presentations here are more closely related to the aspect Process, which may concern any or all of the following topics generally.

- *Measurement*. In particular, there is commonly the issue of dealing with so-called organic data such as text, image.
- *Representation*. This includes the various problems related to over-/ under- coverage, selectivity, domain misclassification, unit problem, etc.
- *Data integration*. When making use of non-survey big data, it is

[1] Li-Chun Zhang (L.Zhang@soton.ac.uk), University of Southampton, UK, and Statistics Norway.

generally necessary to combine multiple sources, not least the statistical population data, in order to achieve the quality that is required of official statistics.

It is helpful to adopt a total error perspective to the processing pipeline (Zhang, 2012; Reid *et al.*, 2017; Rocci *et al.*, 2022), which were initially developed due to the uptake of non-survey administrative sources but are equally applicable to the emerging new sources. Another relevant issue worth attention is the oscillatory tension between stovepipe *vs.* generic standardised solution. Next, regarding data access, it is worth reminding the two quotes below on confidentiality and data minimisation, respectively.

> “Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey [...]” (Skinner, 2009).
>
> “[...] personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed” (GDPR[2]).

The NSI must neither behave nor be perceived as if it were state-sponsored Facebook. Attention is growing on input privacy and proportionality internationally, as in the relevant UNECE and Eurostat initiatives. For instance, Zhang and Haraldsen (2022) develop a conceptual framework and a generic system for secure big data collection and processing, which is planned to be applied for combining retail receipts and debit card payments.

Finally, for the assessment of statistical error, Zhang (2021) proposes the audit sampling inference approach to big-data statistics. Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, auditing aims not to estimate the target parameter itself but some chosen error measure of any given estimator of the target parameter, which may be biased due to failure of the underlying model assumptions or other favourable conditions that are necessary. The framework of inference is design-based given a finite population, from which the random sample is taken under a probability design, while the outcomes of interest and other values known separately from sampling are treated as fixed. Such

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

design-based auditing inference is valid regardless the models or algorithms underlying the estimator being assessed.

In other words, the idea is to utilise the universality of sampling inference for validation, given that the target statistics of acceptable quality can be produced based on non-survey big data directly. It can be appreciated in terms of the long-standing debate between design *vs.* model-based inference approach to survey sampling, as summarised in the table here.

Assessing a design-based estimator with respect to the known sampling distribution yields the design-based approach, whereas assessing a model-based estimator under the assumed model yields the prediction approach.

Inference/Property	Motivation of Estimator	
	Design-based	Model-based
Design-based	Survey sampling	wAudit sampling
Model-based	*e.g.* "weighting is inefficient"	Prediction

Assessing a design-based estimator under an assumed model often leads to claims such as 'weighting is inefficient', whereas assessing a model-based estimator with respect to the known sampling design tend to generate warnings against sensitivity due to unavoidable model misspecification. By the audit sampling inference approach to validation, one could choose from any competing estimators, whether they are originally derived from some assumed models or the known sampling design, according to the preferred error measure with respect to repeated sampling from the given finite population. The philosophical advantage is that one can thereby avoid the epistemological questions such as the existence of a true model or the possibility of infallible learning from the available data. The practical advantage is that audit sampling is generally less resource-demanding than survey sampling.

As some examples, Zhang (2021) applies audit sampling inference to establish the superiority of scanner-data expenditure weights for CPI compared to traditional Expenditure Surveys. Patone and Zhang (2021) applies the approach to Social Media Index (Daas *et al.*, 2015), using the Dutch Consumer Confidence Survey as the audit sample. Bernardini *et al.* (2022) explore efficient audit sampling design for register-based population size estimation.

While one would keep an open mind for future standard of validity in official statistics, auditing does provide the necessary means for the current accepted standard (not least due to the survey sampling tradition).

At Session 3 of the first Workshop on methodologies for Official Statistics, the following papers were discussed:

- "Overview of the Istat activities and open problems" – presented by Mauro Bruno;
- "A Deep Learning Approach to Land Cover Estimation from Satellite Imagery status" – presented by Fabrizio De Fausti.

The discussion was lead by Professor Piet Daas (Statistics Netherlands) and the session terminated with the point of view of the Statistical Production Department of Istat, by Sandro Cruciani who is Head of Directorate for Environmental and Territorial Statistics.

References

Bernardini, A., N. Cibella, and F. Solari. 2022. "A statistical framework for register-based population size estimation". *Technical Report*. Istat Advisory Committee on Statistical Methods 2022 Meeting. Roma, Italy: Istat.

Daas, P.J.H., M.J. Puts, B. Buelens, and P.A.M. van den Hurk. 2015. "Big Data as a Source for Official Statistics". *Journal of Official Statistics - JOS*, Volume 31, Issue 2: 249-262.

Patone, M., and L.-C. Zhang. 2021. "On Two Existing Approaches to Statistical Analysis of Social Media Data". *International Statistical Review,* Volume 89, Issue 1: 54- 71.

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ". *Journal of Official Statistics – JOS*, Volume 33, Issue 2: 477-511.

Rocci, F., R. Varriale, and O. Luzi. 2022. "Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes". *Journal of Official Statistics – JOS*, Volume 38, Issue 2: 533-556.

Skinner, C.J. 2009. “Statistical Disclosure Control for Survey Data”. In Pfeffermann, D., and C.R. Rao (Eds.). *Sample Surveys: Design, Methods and Applications. Volume 29A*. Chapter 15: 381-396. London, U.K.: New Holland Publishers, *Handbook of Statistics*.

Zhang, L.-C. 2021. “Proxy Expenditure Weights for Consumer Price Index: Audit Sampling Inference for Big-Data Statistics”. *Journal of the Royal Statistical Society*, *Series A: Statistics in Society*, Volume 184, Issue 2: 571-588.

Zhang, L.-C., and G., Haraldsen. 2022. “Secure Big Data Collection and Processing: Framework, Means and Opportunities”. *Journal of the Royal Statistical Society*, *Series A: Statistics in Society*, Volume 185, Issue 4: 1541-1559.

Methodologies for big data at Istat: state of the art and open challenges

Mauro Bruno[1], *Monica Scannapieco*[2]

Abstract

In line with the path taken by the European Statistical System, Istat is investing on innovative methods to harness big data sources and to use them to produce new and enriched Official Statistics products. big data sources are not, in general, directly processable with traditional statistical techniques, indeed the nature and structure of such data sources requires the adoption of new data processing methods. This motivates and justifies the growing interest of National Statistical Institutes in Machine Learning and, more generally, Data Science techniques, which represent a (somewhat) new methodological approach to data analysis.

Istat is currently using Data Science techniques in research and innovation projects, e.g. big data experimental statistics. This paper will provide an overview of Istat's more relevant projects in the big data ecosystem. It will focus on two specific big data-based production pipelines, related to the processing of respectively text sources and image sources, thus addressing the variety dimension of big data. Later the velocity big data dimension, which can be exploited to address the need of publishing timely statistics, will be illustrated through a specific project that recently Istat has been investing on. The paper will also highlight the main open challenges, and, in some cases, the preliminary solutions implemented to solve them.

Keywords: Machine Learning, Natural Language Processing, Text Processing, Image Processing, Big Data.

1. Introduction

The Italian National Institute of Statistics (Istat) has been investigating the potential of big data sources for Official Statistics since 2013. As part of the European Statistical System, Istat followed the strategic objectives stated

[1] Mauro Bruno (mbruno@istat.it), Italian National Institute of Statistics - Istat.

[2] Monica Scannapieco (m.scannapieco@acn.gov.it), Italian National Cybersecurity Agency.

in two reference documents, namely: (i) the Scheveningen memorandum[3], stating that European NSOs should investigate the possible use of big data sources to support the production of Official Statistics and (ii) the Bucharest memorandum[4], which indicated the investments necessary to produce big data-based statistics as part of Official Statistics to all effects and purposes.

Within such a strategic framework, Istat is investing on the use of several big data sources, including:

- Web scraped data. The main projects using such data are related to enterprise characteristics, price statistics, job vacancies statistics, and traffic statistics;
- Satellite images for land cover and green areas statistical products;
- Social media data for sentiment analysis and the production of sentiment indexes;
- Scanner data for price statistics;
- AIS (Automatic Identification System) data for vessels traffic;
- Mobile Network Operator data for mobility and tourism statistics;
- Mobile devices sensor data to support social surveys, *e.g.* Time Use Survey (TUS) and Households Budget Survey (HBS).

From a methodological perspective, the investments have mainly addressed: (i) big data sources' heterogeneity, *i.e.* the Variety dimension of big data and (ii) the Velocity dimension of big data. In relation to the first aspect, *i.e.* Variety, dealing with new sources of data like texts or images required the design and development of proper data preparation pipelines, as well as the introduction of machine learning models for data processing. Moreover, accessing external (private) sources required investments on new data access techniques, *e.g.* web scraping and input-privacy-preserving methods. In relation to (ii), *i.e.* Velocity, real time or quasi real time data acquisition methods were necessary, as well as, the exploration of new methods, like *e.g.* machine learning methods to process sensors data.

[3] Scheveningen Memorandum (2013), https://ec.europa.eu/eurostat/documents/13019146/13237859/Scheveningen-memorandum-27-09-13.pdf/2e730cdc-862f-4f27-bb43-2486c30298b6?t=1401195050000.

[4] Bucharest Memorandum (2018), https://ec.europa.eu/eurostat/documents/13019146/13239158/The+Bucharest+Memorandum+on+Trusted+Smart+Statistics+FINAL.pdf/59a1a348-a97c-4803-be45-6140af08e4d7?t=1539760880000.

In this paper, we will describe two pipelines that are becoming the reference standard for projects dealing with textual data and (satellite) images. We will demonstrate such pipelines in three projects, namely: (i) a Web Intelligence project aiming at estimating enterprises characteristics from enterprises' websites, (ii) the Social Mood on Economy Index, computed from Twitter data and (iii) the Land Cover project, producing statistics and maps from Sentinel-2 imagery. Finally, we will describe a new experimental statistic, that will be available by the end of 2022, aiming at improving the timeliness of statistics based on the Eurostat open data on international trade.

2. Dealing with Textual Data

This section deals with textual data as a source for producing: i) estimates of enterprises characteristics, analysing the textual content of enterprises' websites; ii) innovative statistical outputs using Natural Language Processing (NLP) techniques, aiming at estimating a sentiment score of Twitter messages.

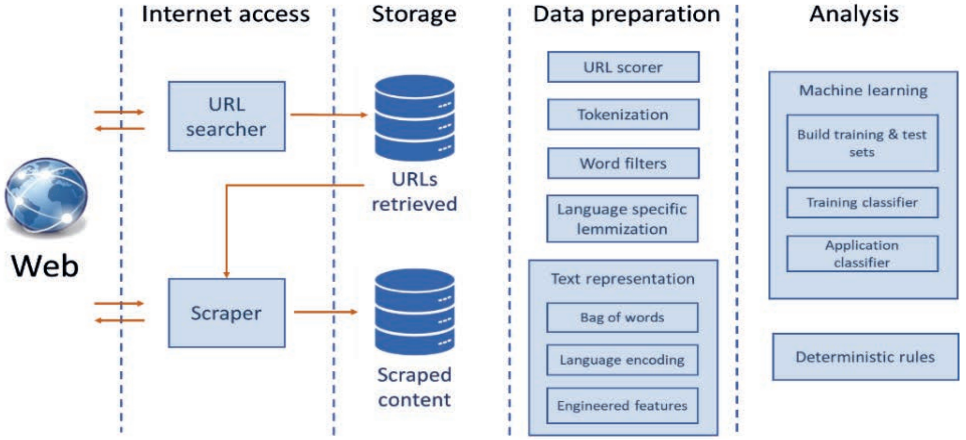
2.1 Enterprise Characteristics Estimates

In 2018, Istat started producing experimental statistics on the activities that enterprises carry out through their websites (web ordering, job vacancy advertisement, link to social media, etc.)[5]. Such activities are a subset of the statistics currently produced by the "Survey on ICT usage and e-Commerce in Enterprises" and are computed starting from enterprise websites' contents, acquired by web scraping tools, and processed with text mining techniques. A machine learning approach is adopted to estimate models with survey data serving as training set for the machine learning task. The textual content of scraped enterprises' websites is used as input to predict the target values by applying the model fitted in the previous step. The experimental statistics are obtained using two different estimators: (i) a full model-based estimator; (ii) an estimator that combines model and survey-based estimates. Considering the various domains for which they have been calculated, the three sets of estimates (survey, model and combined) in most cases are not significantly different (*i.e.* model and combined estimated values lay in the confidence

[5] Detailed information on the experimental statistics are available from https://www.istat.it/en/archive/216641.

intervals of survey estimates). Simulations have demonstrated that the Mean Square Errors of these new estimates are competitive as compared to those produced in the traditional way. The detailed description of the project can be found in (Barcaroli and Scannapieco, 2019).

Figure 2.1 - Generic pipeline for processing textual data from enterprise websites



The *Internet access* phase represented in Figure 2.1 includes two logical blocks, namely: URL searcher and Scraper, while the *Storage* phase includes: i) URLs Retrieved storage block and ii) Scraped content block. To start a collection of data by scraping websites, first a list of URLs identifying the home pages of the sites to be reached is needed. If this is not available, it is necessary to set up a dedicated URL retrieving activity, see (Barcaroli and Scannapieco, 2019) for details.

The *Data preparation* phase includes several blocks related to text transformation needed before the Text Representation can be done this is an essential step when dealing with textual data that are unstructured or partially structured. The Text Representation logical block includes: i) the traditional Bag of Words approach, ii) a language modelling block, *e.g.* recent Word Embeddings approaches and iii) an Engineered Features block, related to the specific features that can be selected for subsequent machine learning tasks.

Finally, the *Analysis* phase for websites is based on the use of a Machine Learning approach, in which either Supervised or Unsupervised Machine

Learning methods can be applied. In addition, for specific use cases, deterministic decision rules can also be put in place: this is the case for instance of a use case aiming at computing if an enterprise is present or not on a social media, which can be checked by looking at the presence on the enterprise website of one of a finite set of social media, *i.e.* Twitter, Facebook, etc.

2.1.1 Accuracy of the results

In our initial works (Barcaroli and Scannapieco, 2019) we performed several tests to assess the accuracy of the machine learning models used to estimate enterprises' characteristics, *e.g. Web Ordering*, *Presence in social media and Job Vacancy*. 2.1 contains the performance indicators of several machine learning models (Logistic Model, Neural Networks, Classification Trees, Naïve Bayes, Support Vector Machine, Bagging, Boosting, Random Forest) used to predict the target variable *Web Ordering*. The model that provided best performance was Random Forest, with an accuracy equal to 0.83 and an F1-measure equal to 0.63.

Table 2.1 - Performance indicators related to the different learners applied for Web Ordering

Learner	Accuracy	Recall	Precision
Naive Bayes	0.84	0.56	0.56
Logistic Model	0.84	0.57	0.57
Classification Tree	0.87	0.64	0.64
Neural Networks	0.88	0.65	0.66
Bagging	0.88	0.66	0.67
SVM	0.90	0.62	0.76
Boosting	0.90	0.71	0.71
Random Fores	0.90	0.73	0.73

Source: Authors' processing

To improve the accuracy of the models, we performed manual inspections of the cases in which survey and machine learning predicted values were conflicting. Once eliminated response errors in the training and test sets, the performance of the different learners resulted to be much higher. In particular, the Random Forest predictor resulted to be very good with an accuracy equal to 0.9 and an F1-measure equal to 0.73 (see Table 2.1), that is, much more than when not considering and treating response errors.

2.1.2 Open challenges

The described project faces (some aspects of) the important issue of integrating a big data source with survey data. In the project, survey data are used as a training set of a Machine Learning classifier executed on Web extracted data. In our initial works (see *e.g.* Barcaroli and Scannapieco, 2019) most quality considerations were made by comparing big data-based estimations with survey-based ones. Recently, in Pratesi *et al.* (2022), more complex data integration methods are used to reduce the bias by combining a probability and a non-probability sample through a vector of common auxiliary variables, as an extension of Kim & Wang, 2019.

2.2 Social Mood on Economy Index

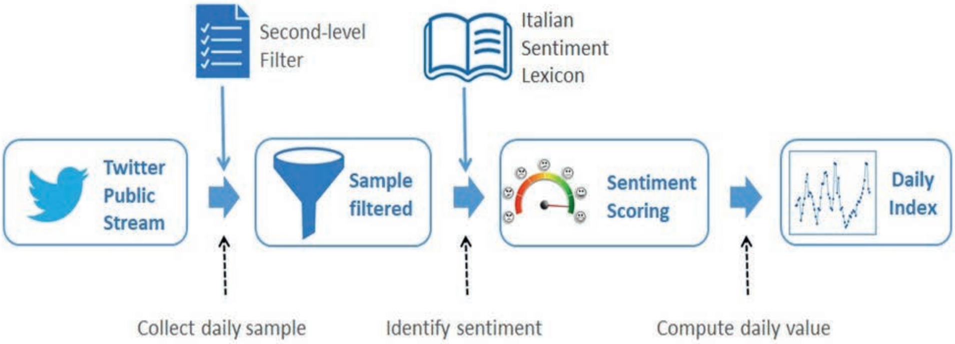
The Social Mood on Economy Index (SMEI) is an experimental statistic published by Istat[6]. It provides daily measures of the Italian sentiment on the economy, these measures derived from samples of public tweets in Italian language captured in real time. The index production procedure collects and processes only tweets containing at least one word belonging to a specific set of filter keywords, which has been designed by subject-matter experts.

The statistical production process handles all the tweets collected in a single day (about 47,000) as a single block. As shown in Figure 2.2, the messages collected, after a cleaning and normalisation activity, undergo a sentiment analysis procedure through an unsupervised lexicon-based approach. The lexicon consists of a set of lemmas associated to a pre-computed sentiment score. Through the comparison between the words of each message (tweet) and the lexicon, a sentiment score is computed and is used to cluster the messages into three mutually exclusive classes: Positive, Negative and Neutral tweets. Lastly, the daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes.

To prevent off-topic tweets from passing the filter and undermining the robustness of the index, we implemented a surveillance system that searches for anomalous values in the time series of the daily index. Daily values

[6] SMEI's experimental statistic is available at: https://www.istat.it/it/archivio/219585.

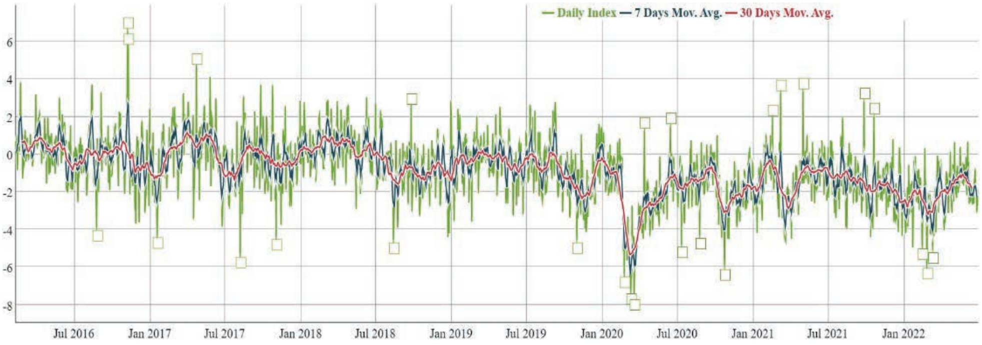
Figure 2.2 - Pipeline to produce the Social Mood on Economy Index



detected as potential outliers generate a set of dedicated diagnostic reports that will be analysed by human reviewers with the aim of deciding whether the detected values are proper data or truly anomalous. All daily index values classified as truly anomalous are imputed via nearest-neighbour interpolation.

Data collection for the Social Mood on Economy Index started in February 2016 and has been active since then almost without interruptions (see Figure 2.3). For organisational reasons, updates of the index are currently published on a quarterly basis. More details on the project can be found in (Zardetto *et al.*, 2019).

Figure 2.3 - Social Mood on Economy index, daily index (green) and moving averages (blue, red)



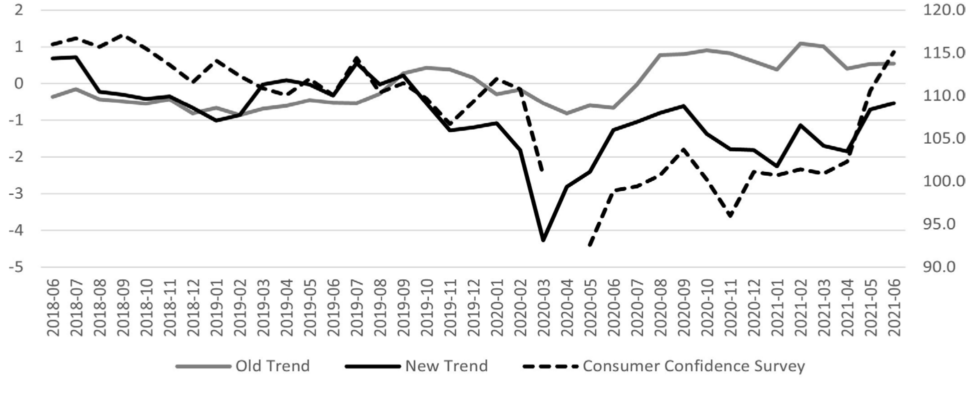
Source: Istat experimental statistics web page on the Social Mood on Economy index (https://www.istat.it/it/archivio/219585)

2.2.1 Accuracy of the results

In a recent work (Catanese *et al.*, 2022), we discussed SMEI's quality issues, and the preliminary solutions implemented to address such issues. In particular, the interpretation of the time series and its evolution in time are highly dependent on design choices, *e.g.* the filter used to select the tweets and the lexicon used to label the words in each tweet. A detailed description of the quality issues can be found in the paper.

In this subsection we show the comparison of 1) the monthly trend of the original SME index (Old Trend in Figure 2.4); 2) the new monthly series resulting from a recent methodological revision involving both the filter and the lexicon (New Trend in Figure 2.4); 3) the Confidence Consumer Index (CCI). The results are very encouraging and prove that the methodological revision increases SMEI's interpretability. Indeed, Figure 2.4 shows that the monthly average of the new disseminated trend series has a better correlation with the CCI. We compare the last three years of the series and for the SMEI trends we use the monthly average value. The Pearson correlation assume value of -0.47 for the previous trend (Old Trend) and 0.71 for the new one (New Trend). The Business and Consumer Confidence Index were not published in April 2022 due to a data collection problem related to Covid and the SMEI was used in that period for the monthly economic note.

Figure 2.4 - Comparison of the previous and the new monthly averages of trends with the Consumer Confidence Survey (monthly data) for the last three years (seasonally adjusted data)



Source: Authors' processing

2.2.2 Open challenges

In relation to SMEI, there are two major research directions that we would like to explore, namely: (i) evaluation of the quality of Twitter's filters and (ii) improvement of the index interpretability.

To evaluate the quality of the filter keywords we have exploited Word Embeddings (WE) methods. To this aim we used *WordEmBox*, an *ad hoc* tool developed by Istat aiming at exploring WE spaces. This software allows to visualise WE models in two dimensions, and to investigate the relationship between words. *WordEmBox* offers three main functions: graph representation, word-analogy, and word-similarity. While word-analogy and word-similarity are the standard test functions to assess the quality of WE model, the graph is the real add-value of this application. It is an original graph-based methodology to explore, analyse and visualise the structure of the embedding spaces.

In the *WordEmBox* is possible to generate and visualise three different types of graphs that we have named: *geometric, linear*, and *geometric oriented*. In the *geometric* graph the exploration range grows quickly, losing the initial semantic focus provided by the seed words. In the *linear* graph, the exploration and the representation are much more focussed, but the model explores just a "narrow" sub-region of the embedding space; the *geometric oriented* graph is a compromise between the previous two.

By conducting human-based interactions with the *WordEmBox*, we can evaluate the quality of the filters in a semi-automatic way. One research activity is the development of appropriate metrics to quantify the goodness of the filter.

To enhance the interpretability of the index and better validate it, we are planning the following future work:fino qui

- Explore the possibility to identify "temporal windows" to correctly evaluate the values of the indicator. Such windows could be identified according to different requirements: (i) understand the change of the volatility characteristics of the index; (ii) understand the impact of event/emergent topics by considering windows centred on extreme values of the index; (iii) explore the "turning" points of the index series. For all the above cases, the interpretability of the index could be improved by looking into the topics/words that are used in the Twitter stream captured. Possible categories of methods that can proficiently be used to the purpose are word embedding and topic analysis.

- SMEI is currently an index based on the computation of a position parameters, namely a mean. It could be interesting to compare it with an index based on dispersion parameters, like *e.g.* variance, coefficient of variation etc.
- Restricting the set of keywords of the filter to focus the attention on more specific economic fields, such as prices/inflation. Evaluate the impact of the filter restriction on the index behaviour and compare the result with similar official statistical outputs.

3. Dealing with images: the Land Cover project

This section deals with a data source type that is very different from the textual data sources described in Section 2. We will describe a project dealing with images and focus on the related pipeline designed and implemented to produce land cover statistics based on satellite images.

3.1 Motivations for Land Cover and Machine Learning

In 2018, Istat launched a project to design and develop an automatic Land Cover (LC) estimation system both to produce statistics and of maps concerning an area and a period of interest. The production of land cover maps and statistics at European level is nowadays performed mainly in two major projects Land Use and Coverage Area frame Survey (LUCAS)[7] and CORINE[8]. Each of these projects disseminates its output on a multi-year basis according to a specific taxonomy, moreover for their production a strong human intervention is required as far as the execution of field measurements, or the photo interpretation of satellite images are concerned. Given these assumptions, a fully automated system providing landcover estimates and maps fed by satellite imagery available for the community today would be

[7] More details on LUCAS project are available from: https://ec.europa.eu/eurostat/web/lucas.

[8] More details on CORINE project are available from: https://land.copernicus.eu/pan-european/corine-land-cover.

both useful and challenging. Indeed, the good spatial and temporal resolution of European sentinel satellite constellations offers the possibility of providing landcover outputs with a temporal frequency hitherto unthinkable.

Machine learning-based classifiers such as Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Decision Tree (DT), and Random Forest (RF) are widely used in the classification of remote sensing imagery, RF and SVM classifiers have shown better performance in terms of overall accuracy (Thanh Noi and Kappas, 2018).

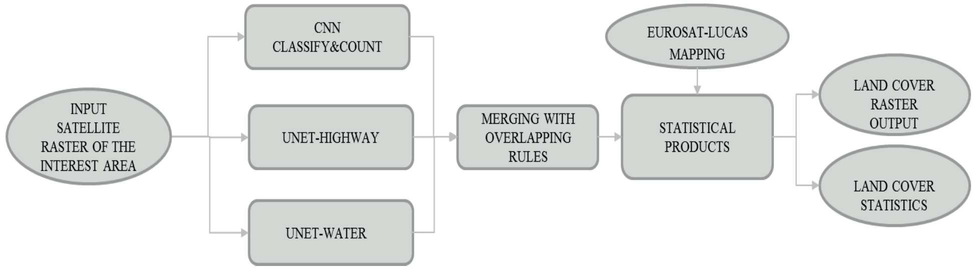
However, these approaches are primarily pixel-based, and exploit the statistical relationship between land cover class and multispectral response as measured by the satellite's on-board instrument. To produce statistics and maps of land cover developed, in Istat we use a different approach that we consider more complete. Indeed, for the classification of land cover in addition to the spectral response of pixel-based approaches we use the spatial patterns that characterise the land cover classes. Making an analogy it is as if for the recognition of objects, we use the colours but also the shapes.

Deep learning algorithms today are the state of the art for spatial pattern recognition and are the ones we have implemented in our pipeline.

3.2 Pipeline for Land Cover Maps and Statistics

In this paragraph, we describe the inference process to produce land cover statistics and maps. We will give just an overview due to the complexity of each individual step; a more detailed explanation of the process can be found in (Zardetto *et al.*, 2021).

Figure 3.1 - Data processing pipeline to produce land cover maps and statistics (inference)



Looking at Figure 3.1, the input raster is composed merging many Sentinel-2 MSI multispectral raster downloaded from Copernicus Open Access Hub[9] relatively to period and area of interest. For each pixel we process mainly the spectral response relative to red, green blue and near infrared bands; for these bands the pixel resolution is 10 metres. Our pipeline can produce land cover statistics for wide areas such as Italian regions.

The input multispectral satellite images are processed by three different deep neural networks trained on different land cover classes, to produce three different output specific of land cover classes.

- *CNN classify and count* is an architecture based on a Convolutional Neural Network (CNN) multi-class classifier INCEPTION-V3. We trained this classifier using an external EUROSAT dataset composed by a collection of Sentinel-2 tiles of 640x640 square metres, labelled with 7 classes: Annual-Crop, Forest, Herbaceous-Vegetation, Industrial, Pasture, Permanent-Crop, Residential. To feed the CNN we decompose the input image through several tiles of dimension 640x640 square metres and compose each prediction to produce the output classification matrix using an original approach we called classify and count.
- *UNET-highway* is an architecture based on a U-net Neural Network (UNET) producing a binary segmentation of raster in input. The output of the classification matrix is the highway land cover class. This architecture shows best performance for identify 1-dimentional land-cover classes such as highway. We trained this UNET network with a dataset that we built for this specific highway segmentation task using as input information Sentinel-2 tiles and highway layer from OpenStreetMap project[10].
- *UNET-water*, we used a UNET to produce a binary segmentation of Water land cover class. We trained this UNET network with a dataset that we built for this specific water segmentation task using as input information Sentinel-2 tiles and water layer from High Resolution Level provided from Copernicus project[11].

[9] https://scihub.copernicus.eu/.

[10] https://www.openstreetmap.org/.

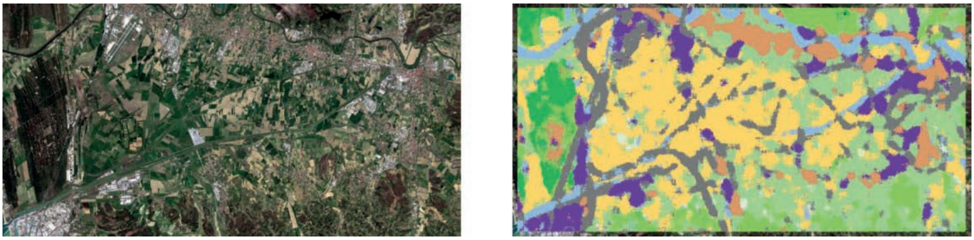
[11] A detailed description of the Copernicus project is available at: https://land.copernicus.eu/pan-european/high-resolution-layers/water-wetness.

In the merging operation the three classification matrixes are merged according to pre-defined priority rules. The class with higher priority is highway, followed by the water class, and then by all other classes. Finally, it is possible to produce outputs based on another land use classification *e.g.* LUCAS classification by applying trans-coding tables.

3.3 Results

In a recent paper (Bernasconi *et al.*, 2022) furtherly investigated the deep learning models described in previous section, to produce automatic land cover maps over large areas such the city of Pisa. The CNN (Inception-v3) has been trained for 120 epochs by means of an NVIDIA Tesla P40 with 32 GB of RAM demanding 110s per each epoch. The accuracy achieved by our model on the validation set is 98.22%. These results are in line with our previous work and with the "state-of-art" literature for deep learning applied to satellite imagery (Helber *et al.*, 2017).

Figure 3.2 - A Sentinel-2 image extracted from top of the atmosphere and around the Pisa city on March the 25th, 2019 (Left). Map produced to calculate the land cover statistics (Right)



Source: Sentinel-2 (https://www.sentinel-hub.com/)

Applying the land cover estimation algorithm to the image in Figure 3.2, authors produced the statistics displayed in Table 3.1. The column "Land Cover Frequency" contains the estimated frequencies for each LUCAS class and the column "Ground Truth" contains the estimates of a ground-truth built by transcoding the CLC project[12] and Lucas. Though not yet satisfying, the

[12] The CLC land cover project is part of the CORINE Programme, it has been designed to provide consistent localised geographical information on the land cover of the 12 Member States of the European Community.

results are quite encouraging, especially for some classes like Crop Land, also considering the possible degree of uncertainty due to the attempt of harmonising classifications.

Table 3.1 - Comparison of land cover statistics produced using deep learning techniques and traditional survey-based estimates (percentage values)

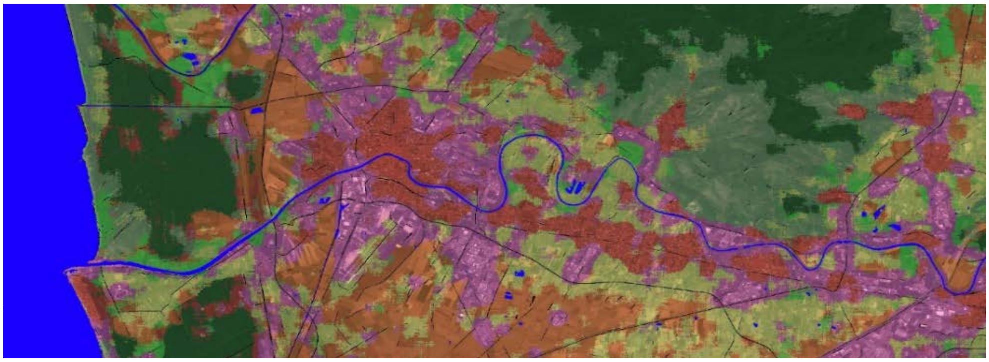
Lucas class	Land Cover Frequency	Ground Truth
Crop Land	51.46	66.96
Artificial Land	29.58	15.58
Grass Land	9.23	-
Water Areas	6.53	2.25
Wood Land	3.16	13.35

Source: Authors' processing

Further, we have tested our system to produce automatic land cover maps over large areas such as the Italian region of Tuscany. Such a processing requires about three days of computation on a server with NVIDIA-V100 GPU. In Figure 3.3 we show a clipping of an area (30120 x 14720 metres) relative to the Arno valley and the city of Pisa.

In this detail we show the good agreement between the satellite image and the land cover map generated by the classification of the three neural networks. We show the good segmentation of the Water class with the Arno River and the well-defined edges. It is also interesting to note the good separation within the urban area between the Residential and Industrial classes.

Figure 3.3 - Land Cover Automatic Map of Pisa Area



Source: Authors' processing

We believe that given the speed of creation of these maps and the good agreement of our system is a useful and promising tool for the analysis and control of the territory and its changes.

3.4 Open challenges

One of the major problems in automated Land Cover (LC) estimation project is the lack of a benchmark to validate the algorithm, according to the chosen resolution and type of classification. Indeed, traditional LC projects use different methodologies and classification and provide different estimates at different resolution that are not easily comparable. Hence the mapping between their classification results (for example between EuroSAT and Lucas, or between Lucas and Corine) is not a trivial task and give rise to several difficulties.

Another difficulty resides in finding a suitable training dataset, compatible with the requested resolution for output. To this aim, we could explore the possibility to integrate input information with administrative sources, like data from regional technical charts, cadastral maps, and agricultural census. Such data are relative only to a given set of locations in the national territory; hence they are not sufficient to build LC maps but can be used to enrich the training dataset to analyse satellite images. To increase the resolution of input information we could exploit also new data from remote sensing like orthophoto or Synthetic Aperture Radar.

Considering the results, we obtained using deep learning algorithms in this domain, we expect that a training dataset enriched in this way could lead to very satisfying outputs.

4. Improving data dissemination timeliness: the new experimental statistics TERRA

This section describes a new experimental statistic that will be launched by the end of 2022, TERRA (imporT ExpoRt netwoRk Analysis). TERRA is an open-source dashboard, that allows the exploratory analysis of Eurostat open data on international trade through dynamic and interactive tools[13]. The system

[13] The source code of TERRA is available on github: https://github.com/istat-methodology/terra.

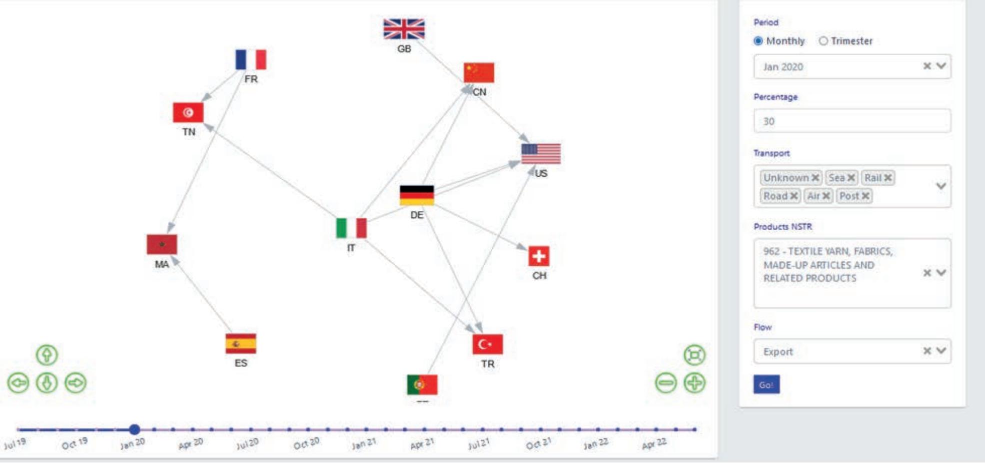
allows to explore phenomena related to the dynamics of global value chains, with the possibility of focussing on specific products and modes of transport.

Implemented in R and Python and based on modern web frameworks, TERRA is one of the first cloud-native applications implemented by Istat.

TERRA monthly processes about one billion records relating to the commercial exchange of goods with foreign countries, produced by the 27 member countries according to harmonised methodologies and publicly available on Eurostat's COMEXT database14. The information base provides the official estimates of trade flows in monetary value and in physical quantities at the maximum granularity in temporal resolution (monthly frequency), characteristics of the traded product, trading partner countries, mode of transport.

The main functions implemented in TERRA allow to analyse the impact of shocks in means of transport and the effects of interruptions in trade relations between countries for specific products with social network analysis techniques offering a set of global indicators, typical of graph analysis.

Figure 4.1 - Graph relating to the export of textile products in January 2020



Period

● Monthly ○ Trimester

Jan 2020

Percentage

30

Transport

Unknown ✕ Sea ✕ Rail ✕ Road ✕ Air ✕ Post ✕

Products NSTR

962 - TEXTILE YARN, FABRICS, MADE-UP ARTICLES AND RELATED PRODUCTS

Flow

Export

Source: Experimental statistics on TERRA (currently being published)

Figure 4.1 shows an example of a network representing the 30% of the global export of Textile Yarn, Fabrics, Made up Articles and Related Product,

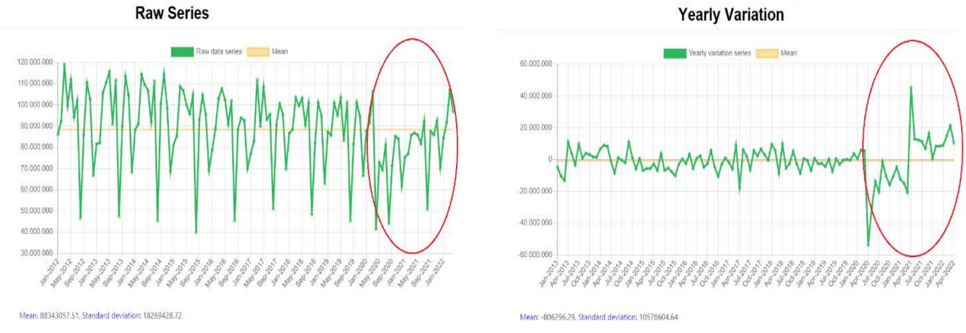
[14] http://epp.eurostat.ec.europa.eu/newxtweb/.

in January 2020. Indeed, the number of countries and trades shown in the graph visualisation depends on the percentage of total trades considered, set through the filter “Percentage”. If we are interested in looking at the graph of all the products’ trades together, it could be necessary to select a low value for the percentage filter, otherwise the resulting graph could be too large and then loose its informative quality.

The dashboard offers another interesting functionality, *time series analysis*. This function allows the visualisation of the raw series and the yearly variations of trade flows between countries in value and in quantity for each product category, while providing statistical tools for the analysis of the series, QQ Norm Plot and autocorrelation function.

As an example, we focus on the export of Textile products from Italy to Germany (see Figure 4.2). We analyse the series in Euro: the plot of the raw series clearly shows that there is seasonality in the path; the export decrease in December and August, when probably Italian factories are closed for holidays, while a peak occurs in October. Looking more in general to the plot it is possible to appreciate a regular dynamic around its own mean, that clearly changes and presents a decreasing structure during the pandemic (highlighted by the red circle), probably due to the lockdown measures introduced by the Italian Government. However, in March 2022 we can see that the level of import is already back to the pre-pandemic period. The seasonal structure changes a lot if we move to the yearly variation: indeed, the plot does not show any regularity and we can appreciate better the effect due to the pandemic.

Figure 4.2 - Export of Textile products from Italy to Germany (Raw series *vs.* Yearly variation)



Source: Experimental statistics on TERRA (currently being published)

Currently, the time span covered by the data relates to months of the last 10 years, the goal is to increase and reach 15 years at the next application update.

4.1 Results

The architectural solution adopted for the implementation of TERRA arises from the need to make the data analysis timely and available to the stakeholders *a few hours* after Eurostat COMEXT monthly data is published. The application retrieves and processes each month data covering a time series of 10 years, managing, and combining 1.375.704 time series representing the level of import and export of 33 Products for each of the 27 European countries versus 192 partners, in Euro or in Kg. Further, TERRA allows to select two different type of monthly time series - raw data and yearly variation, until the latest released update. This corresponds to processing about *1 billion records each month*.

To get and process such a large amount of data, we realised a specific batch programme. The batch script is scheduled to start every 24th day of the month, and, automatically, downloads files from COMEXT portal, performs processing, produces outputs and, finally, updates the data stored on the server.

To speed up the elaboration time, the script runs in cloud platform data is processed using high-performance algorithms, using the library PySpark on Apache Spark Framework. The script is configured to access to COMEXT bulk download section of Eurostat portal and starts a parallel process of downloading 136 files of monthly products data, 2 files for annual products data, 36 files for monthly transport data and the files containing their classifications.

4.2 Open challenges

In its first version, the *time series analysis* section provided a forecasting up to 6 months a-head of the future international trade flows. This functionality used *Google COVID-19 Mobility Reports*[15] to constructs a synthetic indicator capable of explaining the level of restriction imposed in each country using the principal component analysis methodology.

More in detail, the first principal component of Google mobility time

[15] A detailed description of Google COVID-19 Mobility Reports is available at: https://www.google.com/covid19/mobility/.

series was extracted (x_t), *i.e.* the component capable of maximising most of the mobility variability. We then rescaled this data to obtain an index whose values are between zero and one, using the following formula:

$$y_t = \frac{x_t - \min(x_t)}{\max(x_t) - \min(x_t)}$$

We obtained ad new indicator $y_t \in [0,1]$ which assumed a value of 1 if the selected country imposed a total blocking restriction and zero in the opposite case. The new indicator was adopted as a covariate in an interrupted time series model to analyse the effect of the COVID restrictions imposed by each government on the imports and exports of each EU country. However, from October 2022 google mobility data will be not available anymore, therefore data forecasting is not available in the current version of TERRA.

In a future release, TERRA could provide new tools for performing scenario analysis. Indeed, the time series analysis section could be enriched by the inclusion of a subsection dedicated to one or more open indicators such as the Oxford University Stringency Index.

Further, another future goal for the time series is to increase in the volume of processed data and reach 15 years, also searching for new possible leading indicators to be adopted to forecast international trade for all the series considered.

5. Concluding remarks

Big data processing pipelines are a reality in Istat already. The maturity degree of these pipelines and of the related projects is various: some of them are less mature and are related to pilot activities, some others are available as experimental statistics on Istat’s official website having a higher maturity level, some others have already a full maturity and are in production.

Istat is planning to build a full production system based on the use of big data sources, with investments planned for the next years on continuing to build cross-cutting capabilities as well as subject-matter ones for a full-fledged big data-based production. Some of the open issues that need to be faced have been illustrated in the dedicated sections and the very next efforts will be in the direction of working to address them. Surely, the evaluation of the accuracy of the results obtained in the different projects does heavily make use of existing traditional sources, which can have a role either by directly integrating big data sources or by being used as a reference benchmark. While on the accuracy of the projects results, several steps ahead have been made, additional work is needed to refine the current approaches, possibly working in the direction of a generalised reference framework for evaluating accuracy of big data-based products.

References

Barcaroli, G., and M. Scannapieco. 2019. “Integration of ICT Survey data and Internet data from enterprises websites at the Italian National Institute of Statistics”. *Statistical Journal of the IAOS*, Volume 35, Issue 4: 643-656.

Bernasconi, E., F. De Fausti, F. Pugliese, M. Scannapieco, and D. Zardetto. 2022. “Automatic extraction of land cover statistics from satellite imagery by deep learning”. *Statistical Journal of the IAOS*, Volume 38, Issue 1: 183-199.

Catanese, E., M. Scannapieco, M. Bruno, and L. Valentino. 2022. “Natural language processing in official statistics: The social mood on economy index experience”. *Statistical Journal of the IAOS*, Volume 38, Issue 4: 1451-1459.

De Fausti, F., M. De Cubellis, and D. Zardetto. 2018. “Word Embeddings: a Powerful Tool for Innovative Statistics at Istat”. In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data - JADT*, 12th – 15th June 2018: 174-182. Roma, Italy: National Research Council - CNR.

Helber, P., B. Bischke, A. Dengel, and D. Borth. 2019. “EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification”. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Volume 12, Issue 7: 2217 – 2226.

Kim, J.K., and Z. Wang. 2019. “Sampling Techniques for Big Data Analysis”. *International Statistical Review*, Volume 87, Issue S1: 177-191.

Pratesi, M., F. Schirripa Spagnolo, G. Bertarelli, S. Marchetti, M. Scannapieco, N. Salvati, and D. Summa. 2022. “Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics)”. In Balzanella, A., M. Bini, C. Cavicchia, and R.

Verde (Eds.). *Book of the Short Papers. SIS 2022, 51st Scientific Meeting of the Italian Statistical Society:* 305-311. London, UK: Pearson.

Zardetto, D., F. De Fausti, E. Cerasti, A. Pappagallo, and F. Pugliese. 2021. "Deep Learning Segmentation for Improved Land Cover Maps and Estimates". Presentation at *New Techniques and Technologies for Statistics - NTTS 2021*. Brussels, Belgium, 9th - 11th March 2021.

Zardetto, D., C. Fabbri, P. Testa, and L. Valentino. 2019. "New Experimental Statistics at Istat: The Social Mood on Economy Index". Presentation at *New Techniques and Technologies for Statistics - NTTS 2019*. Brussels, Belgium, 11th – 15th March 2019.

A deep learning approach to land cover estimation from satellite imagery

Erika Cerasti[1], *Fabrizio De Fausti*[1], *Angela Pappagallo*[1,]
Francesco Pugliese[1], *Diego Zardetto*[2]

Abstract

Timely and frequently updated Land Cover (LC) information is of primary importance to modern National Statistical Institutes (NSIs). Since LC information needs to be accurate and provided with a high spatial resolution, LC projects are very costly, with very complex production pipelines. Hence, outputs are available at a rather low time frequency. For these reasons, an automatic Land Cover (LC) estimation system using satellite images would be valuable. This paper presents an automated system based on deep learning algorithms to produce LC estimates according to the LC EuroSAT classification. The system takes Sentinel-2 satellite images as input to process data through a convolutional neural network (CNN) and a segmentation neural network (U-Net). Taking the LC EuroSAT dataset as a starting point, a suitable training dataset is built to the scope.

Keywords: Land Cover map, deep learning, remote sensing, pattern recognition

1. Introduction

The possibility to extract Land Cover (LC) information is of crucial importance to modern National Statistical Institutes (NSIs). Since most of the facts and events surveyed by NSIs take place somewhere in the national territory, LC information is structurally complementary to survey and administrative data. High quality LC data and statistics can lead to a wider and deeper understanding of many phenomena of interest.

As far as Europe is concerned, two flagship LC projects exist: CORINE

[1] Erika Cerasti (erika.cerasti@istat.it); Fabrizio De Fausti (defausti@istat.it); Angela Pappagallo (angela.pappagallo@istat.it); Francesco Pugliese (frpuglie@istat.it), Italian National Institute of Statistics - Istat.

[2] Diego Zardetto (dzardetto@worldbank.org, and zardetto@istat.it), World Bank.

(Bossard *et al.* 2000, Büttner *et al.*, 2014), currently run by the Copernicus Programme, and LUCAS (Bettio *et al.* 2002, Eurostat 2003), managed by Eurostat. Despite these projects address the study of land cover very differently – CORINE in a cartography (*i.e.* full-coverage) perspective, LUCAS in a statistical estimation (*i.e.* sample survey) perspective – they suffer common shortcomings. Both are very costly, have very complex production pipelines, rely heavily on clerical work, and produce their outputs with a rather low time frequency. Most of the shortcomings affecting CORINE and LUCAS depend on the huge amount of human workload they require. It is, therefore, very tempting to try to overcome these shortcomings through a process of automation to provide timely and frequently updated Land Cover information.

Given an input satellite image depicting a portion of territory, a fully automatic system should ideally be able to (i) classify the territory according to some standard LC taxonomy, and to (ii) quantify the area (or the proportion) of territory covered by each LC class, without any human intervention.

The Italian National Institute of Statistics (Istat) is currently investigating whether Deep Learning (Goodfellow *et al.*, 2016) methods could be used to derive automated Land Cover estimates of satisfactory quality from Sentinel-2 satellite images. A prototype software system is being developed within the scope of this research.

2. Methodology

An automatic Land Cover (LC) estimation system should be able to take as input a satellite image depicting a portion of territory, and to return as output a table of LC statistics. Although LC estimation is a quantification problem rather than a classification one, we started implementing our system according to a pure ‘classify-and-count’ design (Bernasconi *et al.*, 2022). The main driver of this initial design choice was to incorporate into our system a Convolutional Neural Network (CNN) (LeCun *et al.*, 1989, LeCun *et al.*, 1995), so as to take advantage of its tremendous performance in image classification tasks. Without going into details, our classify-and-count design can be summarised as follows:

1. Train a CNN to predict the LC class of a satellite image 'tile' (*i.e.* a small, fixed-size sub-image).
2. Divide the satellite images covering a 'target area' (*i.e.* the territory for which LC statistics have to be computed) into tiles.
3. Use the trained CNN to predict the LC class of all the tiles generated in step (2).
4. Obtain LC statistics for the target area by computing the absolute or relative frequencies of the predicted LC classes.

We decided to adopt the EuroSAT dataset (Helber *et al.*, 2019) as training set for our CNN. As CNN model, we selected Google's Inception-V3 (Szegedy *et al.*, 2016) architecture, which we customised and trained on the EuroSAT dataset. The adoption of EuroSAT as training set has the following major consequences: (i) our system can only produce LC estimates according to EuroSAT LC classification (only 10 classes, non-hierarchical); (ii) LC predictions can only be obtained for tiles of size 64 x 64 pixels. These points unfortunately prevents any direct comparison among the LC estimates produced by our system and those derived by flagship projects like CORINE and LUCAS, whose LC taxonomies are much richer than EuroSAT one. The analysis of results showed a systematic upward bias affecting narrow linear structures like rivers and highways. This overestimation issue turned out to be inextricably linked to the adoption of EuroSAT as training set and to our tile-based classify-and-count approach. In fact, EuroSAT constrains our CNN to process tiles of size 64 x 64 pixels, whose surface area is quite big (about 41 hectares). But rivers and highways are narrow linear structures. Therefore, any highway fragment framed into a tile occupies just a rather small portion of the tile. Nonetheless, whenever the CNN correctly detects a highway in a tile, our system attributes the whole surface area of the tile to the 'Highway' class. This evidently leads to overestimation. The same happens, of course, when the 'River' class is concerned.

We decided to address the overestimation issue affecting the original design of our system by integrating our tile-based CNN model with a proper image-segmentation algorithm. "Image-segmentation" is the task of localising objects of interest within a digital image (*e.g.* rivers and highways) by identifying their shape and/or contours. The segmentation process groups different image pixels together according to their features, and binds them to a common class label (*e.g.* 'River'). Among several solutions for image-segmentation proposed in the field of Deep Learning, our choice fell on the U-Net model for its intuitive architecture and because it succeeded in many real-world applications. The basic idea that guides the re-design and improvement of our system is simple:

1. Train the U-Net to identify and reconstruct only rivers and highways. Use the trained U-Net to produce partial LC maps of input images (*i.e.* maps where only classes 'River' and 'Highway' are detected).
2. Train our existing Inception-V3 CNN on all the other LC classes of the EuroSAT dataset, and let the old-fashioned classify-and-count approach produce partial LC maps of input images (*i.e.* maps where classes 'River' and 'Highway' are no longer detected).
3. Properly merge the partial LC maps produced in step (1) and (2).
4. For all the pixels of the merged map where the U-Net detected either class 'River' or class 'Highway', trust the U-Net and neglect the predictions of the CNN.

The final output of steps 1. - 4. is an integrated, complete (and still automated) Land Cover map, encompassing all the original EuroSAT LC classes. Automated LC estimates can eventually be obtained from this integrated map by simply computing class frequencies across its pixels.

While the logic underpinning the proposed re-design and improvement of our system is simple, it comes with non-negligible costs in terms of data preparation. Indeed, in order to train the U-Net to perform the segmentation of rivers and highways, a suitable training set has to be constructed for each of the corresponding LC classes. A good segmentation training set for rivers and highways requires (i) samples of diverse Sentinel-2 images representing portions of rivers and highways of different sizes and shapes, along with (ii) the corresponding segmentation masks. These segmentation masks can be thought as binary images (*e.g.* black and white images) where each pixel value encodes the information about the presence (white pixel) or absence (black pixel) of a river in the pixel location. The generation of segmentation masks is, of course, a non-trivial task. We tackled this task by leveraging the auxiliary information on European water areas provided by Copernicus High Resolution Layers and OpenStreetMap's data. In the following we present the data preprocessing for the creation of input images, the procedure for the segmentation masks creation, the U-Net architecture and its application, the resulting regional map and statistics.

2. Data pre-processing and Dataset construction

2.1 Automated generation of raster images of Italian NUTS-2 regions

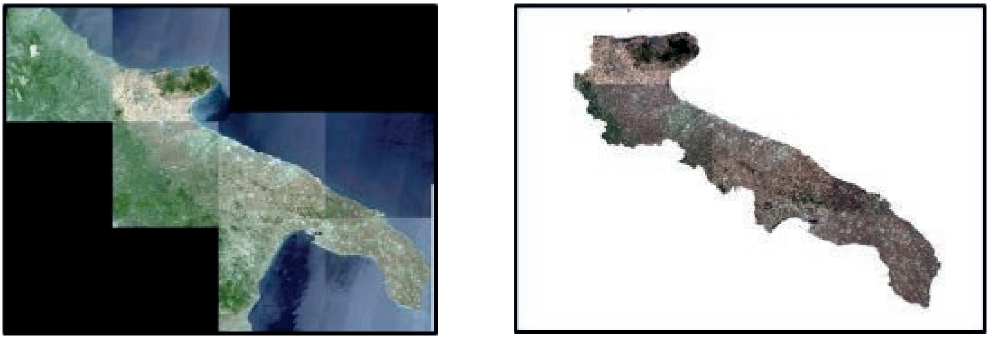
In this Section, we describe the process to build input images for our prototype software system. In order to accomplish this task, a non-negligible data preparation effort is required, as several Sentinel-2 images have to be combined into a geographically consistent mosaic.

With the aim to speed up the generation of a raster image for each Italian region, we implemented a Python script by the GDAL Python library (https://gdal.org/). The script takes as input a shape file containing the administrative boundaries of an Italian region and the set of the collected Sentinel-2 images covering the region. Then, it performs the following steps:

1. Build a geographically consistent mosaic raster image (.jp2) made up of all the collected Sentinel-2 satellite images covering the input region (see Figure 2.1, left panel).
2. Use region's administrative boundaries to clip the region's shape from the raster image built in step 1. Then, save the resulting raster image in a .jp2 file (see Figure 2.1, right panel).
3. Build a "background matrix" whose cells consist of a special artificial texture (see Figure 2.2). The size of the matrix has to match the extension of the mosaic raster image built in step 2.
4. Use the "background matrix" generated in step 3 to fill-up the land or sea area falling outside the boundaries of the target region, within the image built in step 2. The resulting raster image is shown in Figure 2.3.

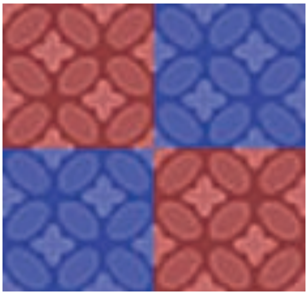
The "background matrix" is intended to enable our classify-and-count Land Cover estimation system to cope with the administrative boundaries of a target region. The "background matrix" is made up of identical tiles, each one comprising 64 x 64 pixels, containing the artificial texture showed in Figure 2.2. We arbitrarily designed this special texture in such a way that, ideally, it could never be confused with any real-world Land Cover class by the classification engine of our prototype software system (that is a customised Inception-V3 CNN).

Figure 2.1 - On the left, mosaic of Apulia Sentinel-2 satellite images. On the right, the same mosaic clipped by the Apulia boundaries



Source: Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu/)

Figure 2.2 - Artificial texture created to identify land or sea areas falling outside the administrative boundaries of a target region. It plays the role of a new 'Other' Land Cover class for our



Source: Authors' processing

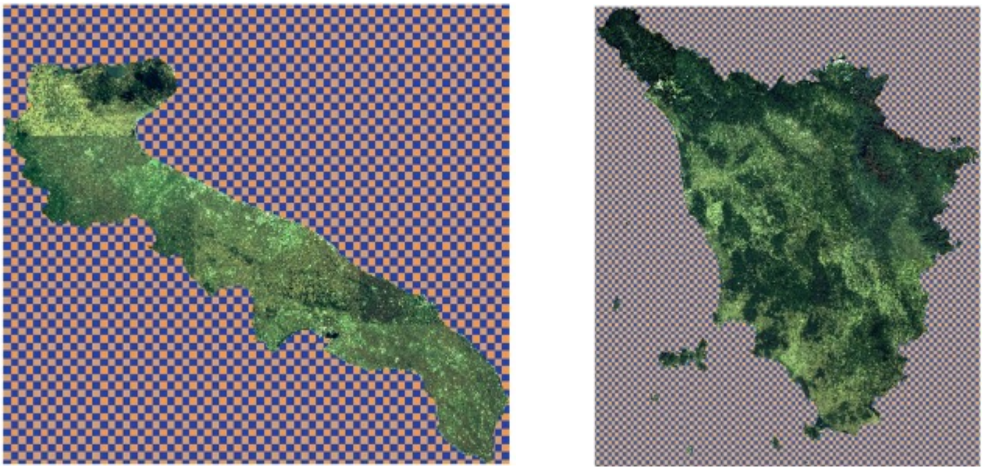
The basic idea behind the "background matrix" can be summarised as follows:

i. Let the texture in Figure 2.2 stand for a new Land Cover class, which we call 'Other'.

ii. Re-train our LC classification engine (*i.e.* the CNN) including also samples belonging to the class 'Other'.

iii. For each tile of the input raster image, generated in step 4, let the re-trained CNN decide whether the tile lies "outside" or "inside" the administrative boundaries, depending on its predicted LC class being equal to 'Other' or not.

Figure 2.3 - Input raster images of two Italian NUTS-2 regions: Apulia (left panel) and Tuscany (right panel)



Source: Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu/)

The output raster images generated for Apulia and Tuscany are shown in Figure 2.3.

2.2 Training set creation: River and Highway classes

A part of the originality and value of this work resides in the creation of a suitable training dataset for the segmentation task. We managed to build a U-Net training set for the 'River' and Highway LC class, as illustrated in this section.

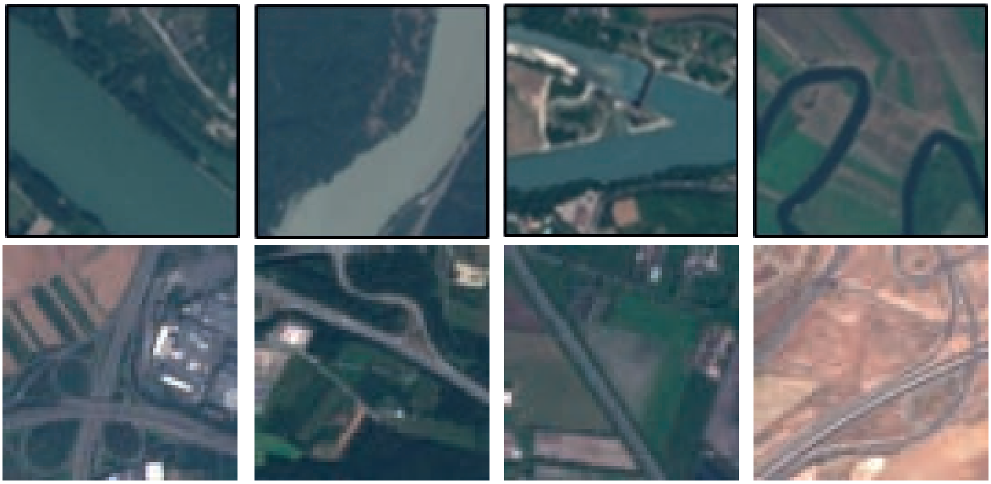
A training set suitable for segmentation tasks of *River* and *Highway* class images requires (i) samples of diverse Sentinel-2 images representing portions of rivers and highways of different sizes and shapes, along with (ii) the corresponding segmentation masks. The segmentation masks are images to be used as label for each pixel of an image: they are binary images (*e.g.* black and white images) where each pixel value conveys the information about the presence (white pixel) or absence (black pixel) of a river in the pixel location (see Figure 2.6).

In the EuroSAT dataset the entire Sentinel-2 image is classified as European river or highway but the information at the single pixel level is not present. The generation of the corresponding segmentation masks is a non-trivial task.

For rivers, we exploited the auxiliary information on European water areas provided by Copernicus High Resolution Layers (https://land.copernicus.eu/pan-european/high-resolution-layers), while for highways we gathered the necessary information from the OpenStreetMap (OSM) project.

The EuroSAT dataset contains 2500 .tiff images belonging to the 'River' class, and 2500 .tiff images belonging to the 'Highways' class, composed by 13 spectral bands (Figure 2.4 offers some examples, where only the RGB bands are rendered).

Figure 2.4 - Example images of the EuroSAT dataset (Sentinel-2 images) for the 'River' LC (upper) and Highways (lower) class

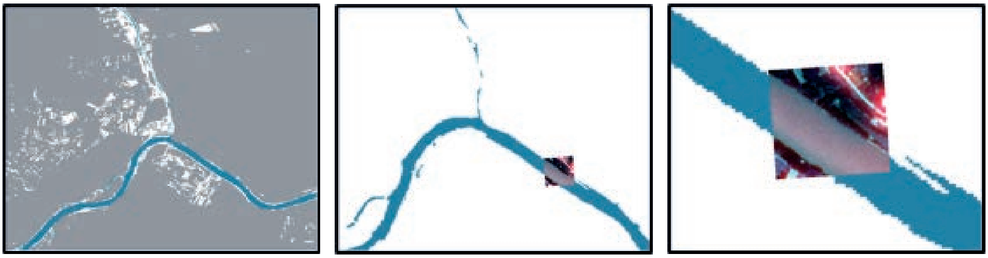


Source: Helber et al., 2019

2.2.1 Rivers

To produce the required segmentation masks we exploited Copernicus "Water & Wetness 2015" High Resolution Layer, which contains data on the following classes: (1) 'permanent water', (2) 'temporary water', (3) 'permanent wetness', and (4) 'temporary wetness'. We decided to use only information about the 'permanent water' class, which is sufficient to map rivers and lakes all over the Europe (see Figure 2.5).

Figure 2.5 - Left: river example derived from Copernicus "Water & Wetness 2015" High Resolution Layer. Middle: overlaying a sample EuroSAT 'River' image on the High Resolution River Layer (QGIS). Right: zoomed version of image in the middle



We used the Python API of the GDAL library (osgeo Python package) to create a script that performs the following steps on each 'River' image in the EuroSAT dataset:

- Take a georeferenced EuroSAT raster image projection.
- Re-project the raster image of the High Resolution River Layer on the EuroSAT image projection (see Figure 2.6).
- Cut the High Resolution River Layer according to the EuroSAT image dimensions, to mask out the portion of High Resolution River Layer not falling in the geographical area covered by the EuroSAT image.
- Read only the information carried by the band 1 (*i.e.* 'permanent water' class) as an array.
- Save the array as a final raster image, after converting in a binary (*i.e.* black and white) RGB image (see Figure 2.6).

2.2.2 Highways

In order to produce the required segmentation masks, we exploited open data from OpenStreetMap (OSM) project (https://www.openstreetmap.org) which offers data on all Europeans routes at different granularity and divided per different categories according to their shape, size, and usage.

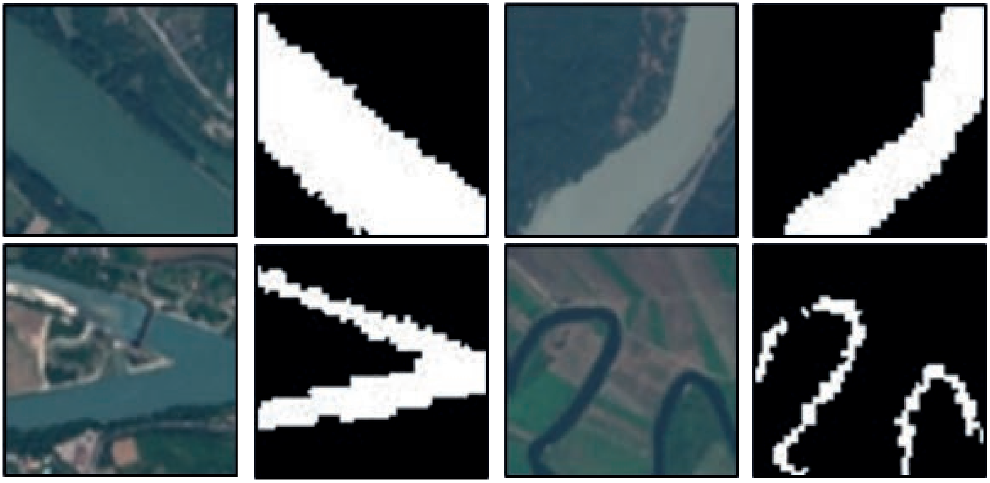
For our scope we are interested in collecting data of routes falling in the category sampled in the LC Highway class. Then, among the provided Map features we focussed on the "Highways" feature category and, after a deep exploration of the data by comparing OpenSreetMap's data with the georeferenced EuroSAT tiles, we decided to use the following categories: "Motorway", "Trunk", "Primary". We downloaded the shape files corresponding to the selected category and put them together in a single file; then we built 2500 "Highways" masks, as described before in the "Rivers" section.

2.2.3 River and Highway class masks

Through these procedures we obtained 2500 segmentation masks for river class and 2500 images for highway class, eligible to represent potential labels for the U-Net training set. In order to get a neat training set, we validated these images by visual inspection and rejected segmentation masks whenever we considered their correspondence to the input EuroSAT image not accurate enough. After the screening, we ended up with 1761 validated segmentation masks univocally associated to EuroSAT 'River' samples and 2496 validated segmentation masks univocally associated to EuroSAT 'Highway' samples. These image-pairs constitute our final U-Net training set for the 'River' segmentation task.

Figure 2.6 below shows two output segmentation masks generated for the EuroSAT rivers images already shown in Figure 2.4, and two output segmentation masks generated for the EuroSAT highways images already shown in Figure 2.4.

Figure 2.6 - The EuroSAT training images of rivers in Figure 2.4 with the corresponding obtained segmentation masks



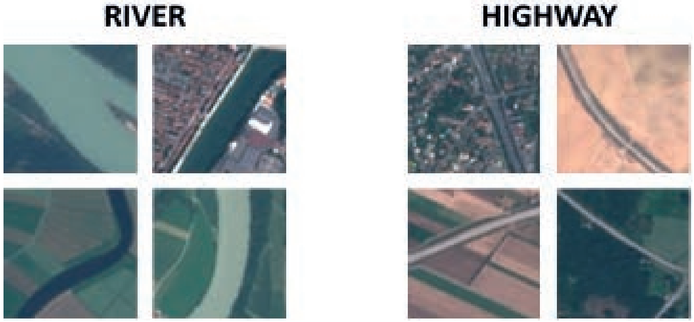
3. Neural network models for classification and segmentation

3.1 CNN classify and count: limits of the approach

As shown in (Bernasconi *et.al.*, 2022), we carefully tested the accuracy of the classify-and-count approach documented therein. We carried out the quality evaluation on sample input images, both by (i) comparing the automated LC estimates produced by our system with available information coming from flagship European projects CORINE and LUCAS, and by (ii) carefully examining visually the corresponding automated LC maps.

This analysis showed that our automated LC estimates had a remarkably good accuracy for most LC classes, except for a systematic upward bias affecting narrow linear structures like rivers and highways. This overestimation issue turned out to be inextricably linked to (i) the EuroSAT training set (see Figure 3.1 for sample EuroSAT patches belonging to the 'River' and 'Highway' LC classes), and to (ii) our tile-based classify-and-count approach.

Figure 3.1 - Sample EuroSAT patches belonging to the 'River' (left panel) and 'Highway' (right panel) LC classes



Source: Huber, 2019

In fact – as explained in (Bernasconi *et.al.*, 2022) – EuroSAT constrains our CNN to process tiles of size 64 x 64 pixels, whose surface area is quite big (about 41 hectares). But rivers and highways are narrow linear structures. Therefore, any highway fragment framed into a tile occupies just a rather small portion of the tile. Nonetheless, whenever the CNN correctly detects a highway in a tile, our system attributes the whole surface area of the tile to the

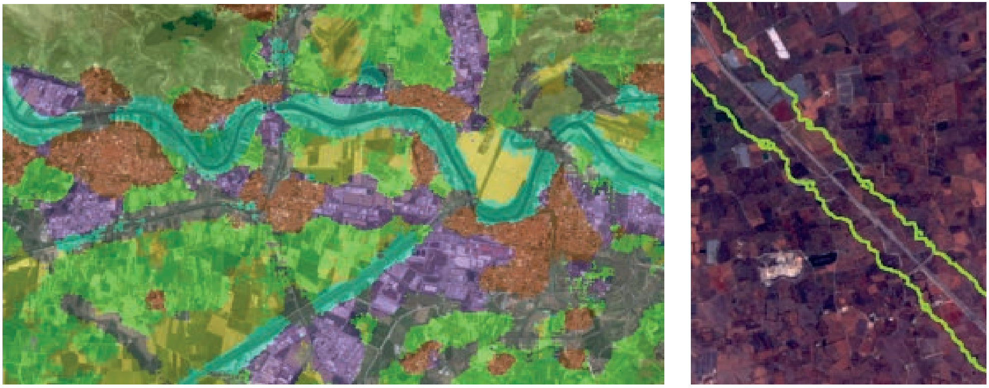
‘Highway’ class. This evidently leads to overestimation. The same happens, of course, when the ‘River’ class is concerned.

Figure 3.2 provides a clear-cut demonstration of the overestimation issue affecting ‘River’ and ‘Highway’ LC classes, and allows appreciating visually the scale of the implied upward bias in LC estimation.

The left panel shows a detailed view of the course of the river Arno, cropped from the north-east quadrant of the ‘Pisa image’ and overlaid with a semitransparent version of the corresponding automated LC map. While the topology of the Arno and of nearby canals seems correctly captured and well described by the map (pale blue areas), the width of the detected water areas is evidently much larger than it should be.

The right panel shows a fragment of a highway cropped from the south-east quadrant of the ‘Lecce image’, along with a green edge-line that outlines the borders of the ‘Highway’ class in the corresponding LC map. To detect this edge-line, the Canny Edge Detector algorithm was used (again, both cited artifacts appearing in the right panel of Figure 3.2 are available in (Bernasconi *et.al.*, 2022)). Unsurprisingly, the average horizontal distance between the green lines, as measured by a GIS (Geographical Information System), almost exactly matches the width of the tiles we exploit in our classify-and-count approach, *i.e.* 640 metres.

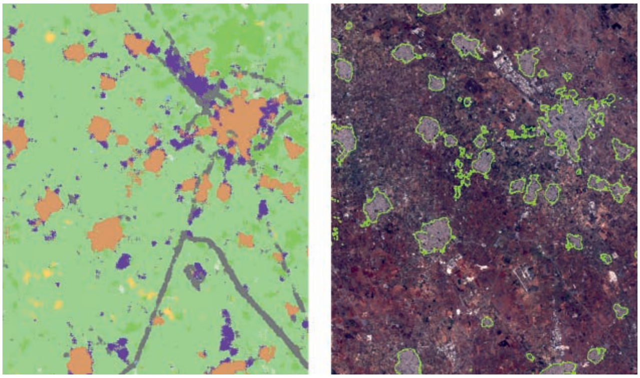
Figure 3.2 - Visual illustration of the overestimation issue affecting our system with respect to LC classes ‘River’ (left panel: bends of river Arno, cropped from the north-east quadrant of the ‘Pisa image’) and ‘Highway’ (right panel: a fragment of a highway cropped from the south-east quadrant of the ‘Lecce image’)



Source: Authors' processing

We stress here that, except for the ‘River’ and ‘Highway’ classes discussed in the present Section, our classify-and-count approach performs remarkably well. For instance, the right panel of Figure 3.3 shows the spatial consistency of our automated LC maps with respect to urban areas. To build this figure, we first extracted the edge of the ‘Residential’ class from the LC map of the ‘Lecce image’ (left panel) using the Canny Edge Detector algorithm. Then, we overlaid the obtained edges on the input ‘Lecce image’ using a GIS. Evidently, the green edge-line outlines with very good accuracy all the cities that are visible in the ‘Lecce image’. Consequently, our automated LC estimates for the ‘Residential’ class do not exhibit any significant (upward or downward) bias.

Figure 3.3 - Left panel: automated LC map of the territory depicted in the ‘Lecce image’. Right panel: The ‘Lecce image’ overlaid with the edge of the ‘Residential’ class (green line) extracted from the automated LC map shown in the left panel



Source: Authors' processing

Coming back to the main weak point of previous system, namely the upward bias affecting ‘River’ and ‘Highway’ LC classes, we decided to address it by integrating our tile-based CNN model with a proper image-segmentation algorithm, namely a U-Net Deep Learning architecture (see next Section for details). The basic idea is simple:

1. Train the U-Net to identify and reconstruct only rivers and highways. Use the trained U-Net to produce partial LC maps of input images (*i.e.* maps where only classes ‘River’ and ‘Highway’ are detected).
2. Train our existing Inception-V3 CNN on all the other LC classes of the EuroSAT dataset, and let the old-fashioned classify-and-count approach

produce partial LC maps of input images (*i.e.* maps where classes ‘River’ and ‘Highway’ are no longer detected).

3. Merge the partial LC maps produced in step 1 and 2 (see next subsections for technical details on how to properly perform this merge operation).
4. For all the pixels of the merged map where the U-Net detected either class ‘River’ or class ‘Highway’, trust the U-Net and neglect the predictions of the CNN.

The final output of steps 1 – 4 is an integrated, complete (and still automated) Land Cover map, encompassing all the original EuroSAT LC classes, plus the ‘Other’ class introduced in Section 2 to take into account administrative boundaries within the input image. Automated LC estimates can eventually be obtained from this integrated map by simply computing class frequencies across its pixels.

3.2 The U-Net model for semantic segmentation

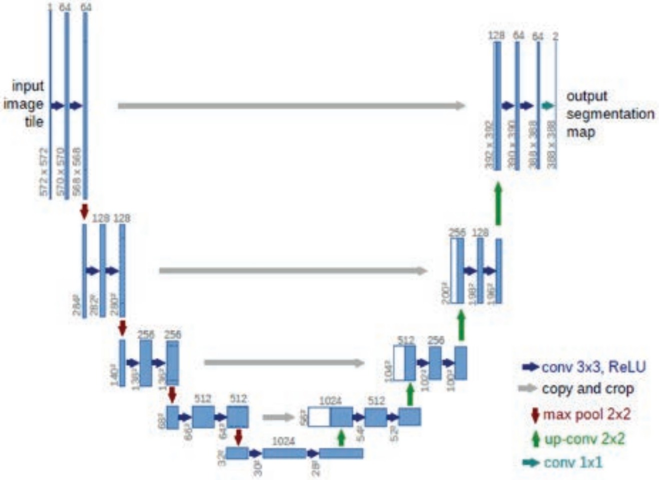
Within the Machine Learning literature, we can identify two types of Segmentation: “Instance Segmentation” and “Semantic Segmentation”. Generally, segmentation refers to the task of localising objects of interest within a digital image (*e.g.* rivers in our application scenario) by identifying their shape and/or contours. Semantic segmentation manages different objects belonging to the same category as a unique entity. On the other hand, Instance Segmentation treats every single object as a different category. Therefore, in order to achieve the highest accuracy, Computer Vision systems should adopt instance segmentation, which lack of supervised datasets. In our work, we focus on Semantic Segmentation process that groups different image pixels together according to their features, and binds them to a common class label (*e.g.* ‘River’). In other terms, Segmentation can be considered as a pixel-level classification, so that it can be very useful for applications aimed to count the number of pixels belonging to objects within the same category. This is what we need in Land Cover estimation.

There exist several semantic segmentations models, we chose to adopt U-Net model (Ronneberger *et.al.*, 2015) thanks to its intuitive architecture and to the fact it was successful in many real applications. The U-Net topology is depicted in Figure 3.4: basically, is a convolutional auto-encoder made of a contracting part able of extract feature (feature extractor) and an expanding part enabling precise localisation and reconstruction of the segmented images. The up-sampling part (*i.e.* the expanding part) repeats rows and columns in order to achieve a large number of feature channels, which enables the network to spread the contextual information to higher resolution layers. Furthermore, a dropout of some entire feature maps (spatial dropout) and a data augmentation (applying elastic deformations on available datasets) allow the U-Net to be trained on a small training set. This increases the U-Net capability to cope with multiple segmentation tasks. As can be seen in Figure 3.4, the contracting section of the U-Net is made up of two 3x3 receptive fields' convolutions. Each convolution is followed by a rectified linear unit (RELU) and a 2x2 max pooling operation for down-sampling. Every down-sampling stage doubles the number of feature channels. The expansive section includes an up-sampling step of the feature channels. This is followed by 2x2 up-convolution that halves the number of feature channels. The final layer is a 1x1 convolution that is used to map the component feature vectors to the required number of classes. Finally, there are skip connections, between pair level layers in the neural network, which feed the output of one layer to another layer skipping a few layers in between. Skip connections foster the spread of information faster in deep neural networks.

In fact, gradient information can get lost as we pass through multiple layers, due to the vanishing gradient issue of the gradient descent algorithm. One advantage of skip connections is that they pass information to lower layers, so that classifying minute details becomes easier. Typically, a stochastic gradient descent algorithm with binary cross entropy as loss function is adopted during the training stage of a U-Net.

The U-Net model requires a supervised segmentation algorithm, which means that labelled samples of already segmented images must be provided to the U-Net during the training phase. The U-Net model must learn how to segment new images into objects of the desired, predefined classes. More specifically, a training set for a U-Net model contains, for each object class, a set of image pairs: one image as example data, and the corresponding segmentation mask as class label.

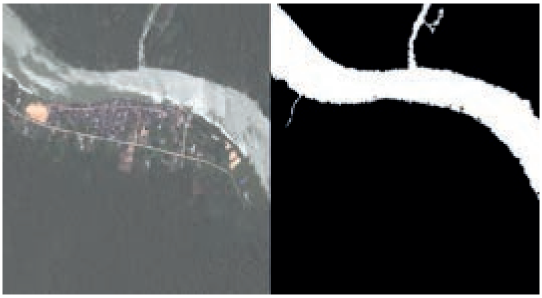
Figure 3.4 - U-Net architecture



3.2 Preliminary U-Net results for the ‘River’ land cover class

In Land Cover estimation from Satellite Imagery, a U-Net can be used to localise the portion of territory occupied by a river as shown in Figure 3.5. The accuracy of the result can be calculated using the Intersection over Union Index (IoU, also known as Jaccard Index) between the predicted mask and a ground truth mask. A score of 0 means complete mismatch, whereas 1 is complete overlap.

Figure 3.5 - Example of river segmentation by a U-Net model

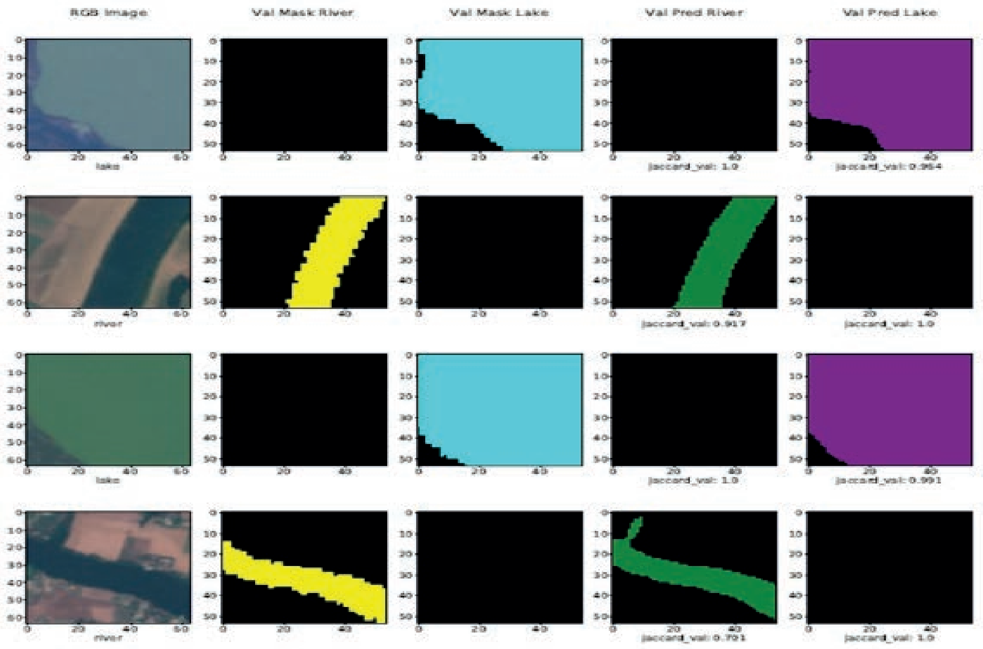


Source: Picture on the left: Huber, 2019. Picture on the right: Authors' processing

In the following, we show the results obtained by the U-Net trained on the ‘River’ segmentation dataset described in the previous Section. The U-Net was trained for 50 epochs achieving, on the validation set, an average Jaccard Index between 0.5 and 0.6. Note that Jaccard scores in this range actually indicate a state-of-art performance, as testified by the best results obtained in Kaggle competitions on similar tasks (see, *e.g.* https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/leaderboard).

Figure 3.6 allows visualising the remarkably good results achieved by the model on a small sample of instances selected from the validation set. Note that in this case a two-class U-Net was used, to show the feasibility of discriminating between different water bodies (*e.g.* rivers and lakes). Indeed, as can be appreciated by inspecting Figure 3.6, our trained U-Net proved able to perform the segmentation of lakes and to correctly discriminate them from rivers. Of course, a single-class U-Net would be enough in applications aimed at identifying water areas without further distinctions, we will name this class waterway.

Figure 3.6 - A sample of U-Net segmentation results. Note that in this case a two-class U-Net was used, to show the feasibility of discriminating between different water bodies (*e.g.* rivers and lakes). Of course, a single-class U-Net would be enough in applications aimed at identifying water areas without further distinctions



Source: Pictures on the left: Huber, 2019. Other pictures: Authors' processing

4. Output and Results

4.2 Merging partial LC maps from U-Net and CNN

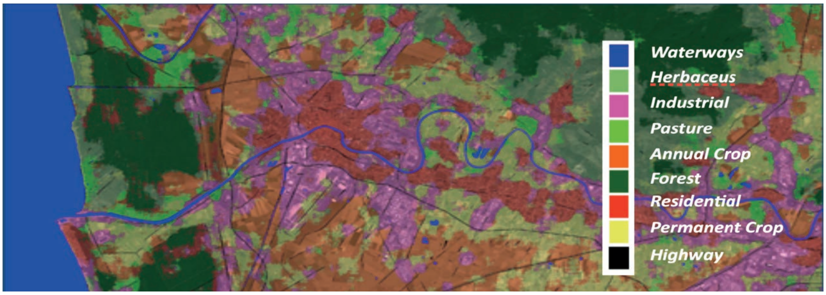
As extensively described before, since classify-and-count approach by CNN suffered an overestimation issue for the particular land cover classes 'River' and 'Highway', a U-Net model was added to pipeline in order to specifically cope with these two "pathological" classes. We use, instead, the Inception-V3 CNN model for the remaining EuroSAT classes ('Annual Crop', 'Forest', 'Herbaceous Vegetation', 'Industrial', 'Pasture', 'Permanent Crop', 'Residential').

To reconcile and integrate the three outputs maps, one produced by Inception-V3 CNN and two by the U-Net (Highway, Waterways), a merge process was devised in charge of producing the complete land cover map from which Land Cover final estimates can be computed.

As a first step of the merge process, a resampling of the Inception-V3 and U-Net output maps to the same spatial resolution is necessary. We use the nearest-neighbour interpolation algorithm for the resampling operation need to be able to not destroy the small areas segmented by the U-Net. The nearest neighbour interpolation is suitable for categorical output maps.

After resampling, partial LC maps derived from U-Net and Inception-V3 can be overlaid cell by cell. The final step assigns to each cell the class identified by the U-Net or the class identified by Inception-V3, according to a hierarchy that depends on the degree of confidence we have on each algorithm. The "Waterways" class must have higher priority over the "highway" class and the "highway" class has priority over all the classes identified by the Inception-V3 Model. An intuition of what has just been said is shown in Figure 4.1 and Figure 4.2, in which we see the result of the merge for the area around the city of Pisa.

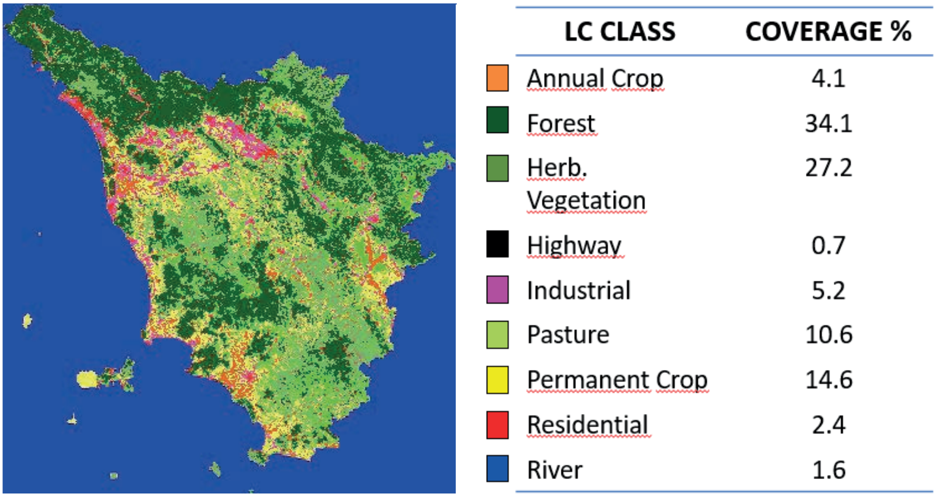
Figure 4.1 - LC Classification of area around Pisa



Source: Authors' processing

Once obtained the merged map we can compute the final regional statistics. Results are:

Figure 4.2 - LC Classification of the Italian region of Tuscany, together with percentage of coverage for each LC class



	LC CLASS	COVERAGE %
🟧	Annual Crop	4.1
🟩	Forest	34.1
🟩	Herb. Vegetation	27.2
⬛	Highway	0.7
🟪	Industrial	5.2
🟩	Pasture	10.6
🟨	Permanent Crop	14.6
🟥	Residential	2.4
🟦	River	1.6

Source: Authors' processing

5. Conclusion

LC statistics can benefit from machine learning methods and algorithms to achieve a high degree of automation in maps and statistics production. This would decrease the effort and improve the frequency of the output. However, several difficulties need to be overcome. Through the creation of a new dataset by integrating information from open data (High resolution layer and OpenStreetMap) we managed to get the LC maps and statistics for an Italian region, Tuscany, by using a pipeline involving two neural networks algorithms for the processing of specific LC classes. Despite encouraging results, we think that in the future we could explore the possibility to attribute the labels using information from administrative sources, like data from regional technical charts, cadastral maps, and agricultural census, to enrich the training dataset. Furthermore, we can increase the resolution of input raster exploring new data from remote sensing like orthophoto or Synthetic Aperture Radar (SAR) by Sentinel1 project.

References

Bernasconi, E., F. De Fausti, F. Pugliese, M. Scannapieco, and D. Zardetto. 2022. "Automatic extraction of land cover statistics from satellite imagery by deep learning". *Statistical Journal of the IAOS*, Volume 38, Issue 1: 183-199

Bettio, M., J. Delincé, P. Bruyas, W. Croi, and G. Eiden. 2002. "Area frame surveys: Aim, Principals and Operational Surveys". In Gallego, J. (Ed.). "Building Agro Environmental Indicators. Focussing on the European area frame survey LUCAS": 12-27. *EUR Report* 20521. Brussels, Belgium: European Commission, Joint Research Centre (DG JRC), Institute for Environment and Sustainability (DG IES).

Bossard, M., J. Feranec, and J. Otahel. 2000. "CORINE Land Cover Technical Guide - Addendum 2000". *Technical Report*, N. 40. Copenhagen, Denmark: European Environmental Agency - EEA.

Büttner, G. 2014. "CORINE Land Cover and Land Cover Change Products". In Manakos, I., and M. Braun (*Eds.*). *Land Use and Land Cover Mapping in Europe. Practices & Trends*: 55-74. Dordrecht, The Netherlands: Springer Nature, *Remote Sensing and Digital Image Processing.*

Eurostat. 2003. "The Lucas survey. European statisticians monitor territory (Updated edition - June 2003)". *Working papers and Studies*. Luxembourg: Office for Official Publications of the European Communities.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning.* Cambridge, MA, U.S.: MIT Press.

Helber, P., B. Bischke, A. Dengel, and D. Borth. 2019. "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification". *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Volume 12, Issue 7: 2217 – 2226.

LeCun, Y., and Y. Bengio. 1995. "Convolutional Networks for Images, Speech, and Time-Series". In Arbib, M.A. (*Ed*.). *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, U.S.: MIT Press.

LeCun, Y., B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition". *Neural Computation*, Volume 1, Issue 4: 541-551.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In Navab, N., J. Hornegger, W.M. Wells, and A.F. Frangi (*Eds*.). "Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015": 234-241. *MICCAI Lecture Notes in Computer Science, Volume* 9351. Cham, Switzerland: Springer.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. "Rethinking the Inception Architecture for Computer Vision". In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Proceedings*. 27th – 30th June 2016, Las Vegas, NV, U.S.: 2818-2826.

SESSION 4

Standardisation of methods and processes

INTRODUCTION

Maurizio Lenzerini[1]

In 2016, the Italian National Institute of Statistics (Istat) has begun a Modernisation Programme as a strategy to improve the current statistical offer and move towards more modern means of data production (Vaccari *et al.* 2022). Such a strategy implied the removal of the obstacles that hindered the harmonisation of statistical processes causing redundancies and bottlenecks. In particular, to increase the effectiveness and the efficiency of the production chain, the reengineering and the standardisation of the business statistics processes have become of primary importance. Traditionally, business surveys use customised methods and tools, with the risk that statistical processes become tied to specific persons, relying on their knowledge and skills. This organisation produces potentially dangerous effects, such as duplicated work and limited reuse of tools and competencies (Bruno *et al.* 2018). The Generalised Process for Business Statistics (GPBS) project aims to identify and implement a general data model and architecture, with the goal of standardising similar steps of business surveys. More precisely, coherently with the Generic Statistical Business Process Model (GSBPM) framework, the GPBS initiative has the main objective to standardise:

- Methods and tools for the statistical phases following the data collection stage.
- Specification of workflow, meant as the most appropriate combination of data and methods to be used in the statistical process.
- Metadata, in order to harmonise concepts and data structures at domain level.

As for the first two items above, Istat has established a specific Working Group on methods and standards (WG), with the goal of developing a catalogue of methods, tools and statistical services and to extend the set of standard methods and tools for statistical production, so as to implement statistical services according to the Common Statistical Production Architecture (Vaccari et al. 2022). As for metadata, official statistics is making increasing use of registers, which collect and integrate data from both administrative

[1] Maurizio Lenzerini (lenzerini@diag.uniroma1.it), Sapienza Università di Roma.

and statistical sources to allow the production of a wide range of consistent statistics. Having registers as a basis, new and appropriate statistical processes can be put in place. A new, layered metadata architecture integrated with the registers' setting may help greatly with respect to different purposes, such as:

- Allowing proper statistical analyses on registers' data by providing an explicit and well-defined statistical semantics to microdata present in registers.
- Deriving output statistical datasets from statistical data cubes that, in this case as well, have an explicit and well-defined statistical semantics.
- Governing the whole production pipeline from registers' data to output datasets by means of a defined and coherent metadata asset.

Given its importance and relevance within Istat, several papers on standardisation of methods and processes have been presented to the Comitato Consultivo per le Metodologie Statistiche (Advisory Committee on Statistical Methods) of Istat. The Advisory Committee participated in several interesting and stimulating discussions and provided advice and suggestions for improving the approach used in Istat. The following is a list of the various papers presented to the meetings of the Advisory Committee:

2017

- "On the design and implementation of a Generalised Process for Business Statistics", by Bruno, M., D. Infante, G. Ruocco, and M. Scannapieco.
- "The Italian Integrated System of Statistical Registers: design and implementation of an ontology-based data integration architecture", by Radini, R., M. Scannapieco, and L. Tosco.

2018

- "A new framework for quality assessment of processes based on Integrated Administrative Data", by Rocci, F., R. Varriale, G. Brancato, and O. Luzi.

2021

- "Longitudinal and cross-sectional analyses of data in the Integrated System of Statistical Register", by Altarocca, F., M.R. Aracri, R. Benedetti, R. Radini, and G. Vaste.

2022

- "On designing aggregated data as Statistical Data Cubes", by Scannapieco, M., M. Scanu, L. Tosco, A. Bianco, and M.K. Riccio.

At Session 4 of the first Workshop on Methodologies for Official Statistics, the following papers were presented:

- "Standardisation of methods and processes: Overview of the Istat activities and open problems" – presented by Carlo Vaccari;
- "Metadata for statistical processes on registers" – presented by Mauro Scanu.

The discussion was lead by Mr Fabio Ricciato (Eurostat) and the session terminated with the point of view of the Statistical Production Department of Istat, by Alessandro Faramondi, head of Division for structural statistics on businesses, governmental and non-profit organisations.

References

Bruno, M., G. Ruocco, M. Scannapieco, and O. Luzi. 2018. "Standardization of Business Statistics Processes in Istat". Paper presented at the *European Conference on Quality in Official Statistics - Q2018*. Kraków, Poland, 26th - 29th June 2018.

Ascari, G., M. Ballin, S. De Francisci, and C. Vaccari. 2022. "Standardization of methods and processes: Overview of the Istat activities and open problems". Presentation at *Istat's Workshop on Methodologies for Official Statistics*. Roma, Italy, 5th - 6th December 2022. https://www.istat.it/en/archivio/277812.

Standardisation of methods and processes: overview of the Istat activities and open problems

Gabriele Ascari[1], Marco Ballin[1], Stefano De Francisci[1], Carlo Vaccari[1]

Abstract

This paper aims to offer an overview of the activities carried out by Istat towards the standardisation of statistical processes, with a focus on the tasks implemented within the Directorate for Methodology and the Working Group on standardisation of methods and tools. Standardisation of processes, methods and tools is an important step towards the efficiency of statistical production. The activities of the Working Group have been built on well-known international models, such as GSIM and, GSBPM. The opportunities offered by the proposed process analyses and the potential future developments are described.

Keywords: Process standardisation, Statistical processes description, GSBPM, GSIM.

1. Introduction

Standardisation of statistical processes is a key element towards a more efficient statistical production, which in turn is necessary to guarantee benefits both for producer and users of official statistics, for example in terms of more timely data releases, improvements in cost efficiency and, more in general, for a better quality of statistics.

In 2016 The Italian National Institute of Statistics (Istat) launched a Modernisation Programme as a strategy to improve the current statistical offer and move towards more modern means of data production (Istat, 2016). Such a strategy implied the removal of the obstacles that hindered the harmonisation of statistical processes causing redundancies and bottlenecks.

[1] Gabriele Ascari (gabascari@istat.it); Marco Ballin (ballin@istat.it); Stefano De Francisci; Carlo Vaccari, Italian National Institute of Statistics – Istat.

Within the Modernisation Programme was a reform in the organisational structure, with the separation between a production department and a support department. This implies, among other things, that efficiency in statistical processes derives from successful communication/collaboration between the two areas. This can be easily achieved if the services offered by support teams are clearly expressed and are supported by tools that implement high-standard methodologies for each step of the process production.

Therefore, the standardisation of statistical processes is strictly related to the standardisation of services which are at the basis of Istat's Modernisation Programme and modern official statistical production. Before addressing the main activities and solutions explored by Istat for this purpose, it may be useful to highlight that such solutions should be based on the most well-known and valid standardisation models used in modern statistics.

For the scope of this paper, we will often address the GSBPM and GSIM models (UNECE 2019, 2021). The first is an international standard developed by UNECE which has been adopted worldwide to describe statistical processes; the latter is an accompanying tool used to describe statistical information through a set of standardised objects and the identification of their relationships. These models belong to the family of ModernStat models (GSBPM, GSIM, GAMSO, CSPA) which are currently used, at different levels of adoption, by most official statistics producers.

In section 2 the need of standardisation determined by new Istat business architecture is described. Section 3 will focus on the tasks and main goals of the Methodology working group, especially the harmonised description of some process and its documentation.

Conclusions, some notes on the future developments and on catalogue of methods and tools are contained in last section.

2. Standardisation at Istat

As mentioned in the introduction, modernisation and standardisation are related concepts. The Modernisation Programme explicitly mentions that the new Istat Business Architecture should privilege the harmonisation and standardisation of procedures. In such programme is mentioned the development of a service catalogue that should identify the available services, which can be of different natures. The focus of this work will be on methodological services (that is, the ones developed and offered by the Directorate of Methodology and Process Design), but it should always imply that there are many other services both inside and outside the technical-scientific category.

SINTESI project (Sistema INTegrato per le Statistiche sulle Imprese) can be seen as an example and a first step towards the standardisation of both processes and services in a specific statistical domain. Its purpose was to address the stovepipe architecture of the enterprises surveys and move towards a structure based on a web platform which allowed for a horizontal integration of the processes and the use of centralised assets for the collection, process and analysis of the data.

The ideas and the objectives behind this project were at the basis of an initiative of much greater scope: the Working Group on methods and standards (WG). This WG, which was internal to Istat’s organisation and specific to the Directorate for Methodology, carried out its activities for 18 months beginning near the end of 2019. Its main aims were: "to develop a catalogue of methods, statistical tools and services", "broaden the set of standard methodologies and tools for statistical production", "create statistical services according to the CSPA architecture". In other words, the objective of the WG was to extend the objectives of the SINTESI project to all services involving methodology, in order to offer a catalogue that can be easily used by any production sector

3. Activities of the working group

3.1 Standardisation of processes description

The attention of the WG focussed on the following processes supported by the methodological Directorate: sampling design, editing and imputation, confidentiality protection, big data activities, small area estimation.

It is clear that they are not the only processes that need methodological support. Such choice has been the result of a pragmatic approach that search for the balance between the services supplied and the resources available for the WG activities.

Given the set of processes that need to be described, one of the first difficulties encoutered by the WG arose from the necessary involvement of many members of the methodological direction (one team for each process supported). Each of these teams has a very in-depth understanding of the methodologies and process.

It happens that a deep specific knowledge match with the adoption of a “methodological slang” (or “jargon”) and with a different feeling about the details needed for an appropriate description of process step.

To overcome the “jargon” and to produce some harmonised descriptions of the process step it was necessary to adopt a common language between the teams members, so that the concepts described could be easily understood without ambiguity outside each specific group.

As stated at the beginning of this document, GSBPM and GSIM models were the natural choices to model the description of the processes and services.

Nevertheless, the knowledge of such international standards (in particular GSIM) are not so widespread even between methodologists and the terminology of such standards is not the one used in the daily work by statisticians in Istat.

Since it was not possible to offer courses on GSIM to all people involved in the project, in this first phase it was necessary to adopt an intermediate language between that of GSIM and that commonly used.

The other difficulty mentioned and arising from a deep and specific knowledge developed by each expert, concerns the detail level (or the “granularity”) that should be used in the descriptions of the processes. As for the case of the terminology, in this phase was adopted as a common “standard” between the teams having in mind the GSIM standard.

In both cases, terminology and granularity, the final balance adopted was the result of some trials carried out with the support of some “facilitators” operating across the specific teams.

Once each team finished with the description of their process, a cross-comparison between teams has been carried out. The description of each process has been assigned to another group with the task of reading it and identifying the points to be clarified, those that needed further specifications as well as completeness considerations about the flow description.

The resulting filled forms, exemplified in annexed Table 1, have been the main input for the steps described in the next paragraphs. Such Table 1 contains the first seven rows of the fifty used to describe the GSBPM "5.1 Integrate data" phase.

Moving from left to right of the Table 1 we can observe an increase of the details and in most cases a very short description of action is accompanied by a "narrative description" of the objective pursued in such row.

For example, the first two rows concern the data acquisition of the sources that must be integrated. As it can be seen, the short description of the action "Reading of the first source" is followed by "narrative description". For this row, such narrative description is quite simple "The data of the first source are loaded along with the names of the existing variables" but in most cases it can contain more details (see the following rows of the same Table). In following columns some information about needed metadata "Metadata for reading (path, file name, record layout, psw., etc.)" are reported. Indeed, the central part of the Table, column-wise, is dedicated to the description of the GSIM objects that are relevant to each step including the inputs and outputs of each step. Thus, following the example being described, the metadata used for the reading step can be considered as a "Process support input", while the input itself that is being transformed are the data from the sources that have been acquired. As for any step, such input will be transformed into something else, which can be obviously identified as the output: in this case, the datasets that will be available for the next steps of the process.

The last columns of the Table show the methodological tools used for each step, mainly the statistical methods and procedures (for example the R or SAS functions).

Taking a step back to the GSIM objects, it is important to note that each input and output is identified by custom names: for example, in Table 1 the two sources to be integrated are labelled input 1 and 2. This allows the reader to follow the 'journey' of each piece of data or information as it passes and is transformed through different procedures and steps. For the integration process, the data acquired and stored in memory in the first step are used, for example, as the input to compute quality indicators during the macro step of the analysis of the variables.

In summary, the forms offer multiple views on the main methodological processes:

- a vertical step-by-step view of the process (processes flow);
- a horizontal step-specific description (input and tools need to carry out each step with a short description of the objectives and output produced);
- a diagonal view that followed the path of the inputs and outputs of a process.

3.1.1. Some notes about the filled forms

Although each team did not need to fill in the form to carry out its activity, it was widely recognised by them that the effort made for the description of the production flow required to focus attention also on steps that for team's specialists did not represent elements of great importance, but which were found to be fundamental in making them understandable to specialists in other sectors (it is one of the output of the cross check between teams).

Since the forms have been designed and filled in trying to guarantee their readability also to non-experts in the specific sector, we believe that they can be used as a tool for dialogue with production services and consequently to better fit the support to the needs and the role of each department.

Although each form has its great importance even individually, the real added value is the availability of the same information about different statistical processes. The whole set of forms help in having a "measure" of some problems cutting cross the services supported.

For example, the set of forms highlight as the ETL (Extraction, Transformation and Loading) processes represent a significant percentage of daily operations in statistical processes. In some cases, it underlines as some processes are heavily based on *ad hoc* procedures (it means procedure developed with a view on a too specific problem and stocked on personal PC). In other cases, it shows as some "methodological standards" need to be updated.

Reading the forms of different processes can help to compare methods and tools used, find bottlenecks and best practices, helping a process of standardisation between processes

Summarising, the set of filled forms can help in defining and sharing the priorities for research and for the need of development of application tools and web services too.

Each form, if applied to a specific survey, should be considered both as a check list of the information that have to be supplied to the team in charge of the service support and once it has been filled in, it stores the information and the choice concerning the tools or the statistical methods adopted by the survey.

The forms can be also easily used as a basis for the construction of on-the-job training courses.

3.2 The ‘Collection of methods’ website

For sake of usability, it was considered that the wealth of information accrued for the purpose of process description could be better arranged on a dedicated platform. The idea was to build an easy tool offering to the user an introduction to the set of services made available by the Methodological Direction and to the supporting teams. This tool should be easily updated every time a new method or procedure is made available for the production processes.

The chosen infrastructure and solution have to be considered as an intermediate step towards the development of a new catalogue (up to now a pilot version of the new catalogue has been developed).

The present solution is a user-friendly website built with wiki tool made available by Sharepoint, the content management system standard within the Institute, so that the also the look-and-feel of the new website would be familiar to the production units involved.

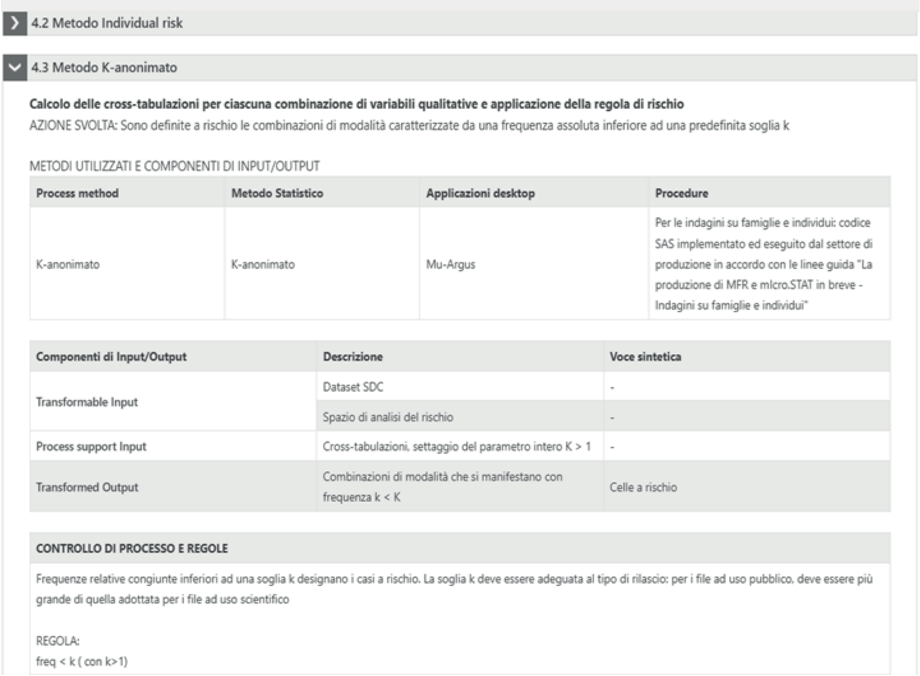
The website follows the scheme used for the process documentation activities, and therefore it heavily relies on the concepts used by the GSBPM and GSIM models. Indeed, the homepage explains the purpose of the website and its structure, introducing the visitor to those concepts. The main areas are the ones studied in the working group and for some of them some specific sub-areas are introduced (for example, for the sampling business function, both sampling for economic surveys and for social surveys are described).

Each area (process) or sub-area has its own page where the main steps of the methodological process are described. The website layout is flexible so that the user can click and expand only the sections or paragraphs they are interested in. The information that can be found, however, is the same for each section and includes a description of the main purpose of the methodological process, the implemented actions, the input and output methods that are used, the input/output components, and, in some cases, a simplified version of the code used for the method, usually in R or SAS (Figure 3.1). A glossary and a list of the main software used will complete the website. The website is currently available only in Italian.

In Figure 3.1 a section of the Confidentiality methodological area is shown, concerning specifically the sub-area “microdata”. The macro step “Risk evaluation” is made up of different steps that can be expanded by the user. Each step, like the one shown in the Figure, includes a description of the action and the concepts mentioned in paragraph 3.1, such as the process method, the statistical method or the desktop applications. Furthermore, the GSIM concepts used in the process description (transformable input, process support input, transformed output) are also included.

Obviously, one of the main difficulties with such a platform with a large and growing availability of information is its maintenance: this will be a challenging task from both a methodological and a technical perspective. Indeed, the website can be considered as a “photograph” of the state of the offering of methodological service for a specific period; but as methods are updated, discovered or made obsolete the need will arise to keep track of such changes. At the same time, the website will have to follow the changes of the methodological catalogue but, more importantly, it will have to be interlinked to it. This is the most important feedback received during the development of the website and this need will have to be addressed while the website and the future catalogue will undergo the appropriate refinements.

Figure 3.1 – Example from the web page of the Confidentiality area – microdata



4.2 Metodo Individual risk

4.3 Metodo K-anonimato

Calcolo delle cross-tabulazioni per ciascuna combinazione di variabili qualitative e applicazione della regola di rischio

AZIONE SVOLTA: Sono definite a rischio le combinazioni di modalità caratterizzate da una frequenza assoluta inferiore ad una predefinita soglia k

METODI UTILIZZATI E COMPONENTI DI INPUT/OUTPUT

Process method	Metodo Statistico	Applicazioni desktop	Procedure
K-anonimato	K-anonimato	Mu-Argus	Per le indagini su famiglie e individui: codice SAS implementato ed eseguito dal settore di produzione in accordo con le linee guida "La produzione di MFR e mIcro.STAT in breve - Indagini su famiglie e individui"

Componenti di Input/Output	Descrizione	Voce sintetica
Transformable Input	Dataset SDC	-
	Spazio di analisi del rischio	-
Process support Input	Cross-tabulazioni, settaggio del parametro intero K > 1	-
Transformed Output	Combinazioni di modalità che si manifestano con frequenza k < K	Celle a rischio

CONTROLLO DI PROCESSO E REGOLE
Frequenze relative congiunte inferiori ad una soglia k designano i casi a rischio. La soglia k deve essere adeguata al tipo di rilascio: per i file ad uso pubblico, deve essere più grande di quella adottata per i file ad uso scientifico REGOLA: freq < k (con k>1)

Source: Collection of methods website

3.3 Process documentation

The acquisition of information about methodological processes has allowed a set of patterns to emerge and, although some of them were already known to researchers within the Directorate, it has been possible to identify the issues and the potential solutions in more specific ways.

As it was previously mentioned, one of the main elements in this regard was the large number of *ad hoc* procedures dealing with input acquisition (“reading” procedures). Almost every methodological area has implemented its own way to deal with this task, which includes data acquisition, data conversion and dataset merging functions. The variety of such functions, often implemented in different programming languages, is related to the habit of each sector and the need to deal with specific data formats that can be different among different areas. However, the analysis carried out for the description of the methodological business functions should be further improved to check the possible presence of common characteristics in the data acquisition procedures, so that a harmonisation process can be carried out at least for these similar functions and the reuse of the same code could allow for improved efficiency.

Of course, a similar approach could be adopted for other functions such as diagnostic procedures, parameter estimations and so on, which probably have some common underlying theme that could allow the development of standardised routines for such tasks. In the following, it is suggested as the same description could, in principle, to be used to draft the methodological notes that are published together with the data. Up to now, these notes are obviously designed for users who need a description (more or less detailed) for an informed use of the data issued.

Currently, the notes published by Istat only partially follow a standard scheme; although the product layout is easily recognizable, the content of some sections does not follow strict requirements.

It is possible to assume that the granularity used for the description of the processes and the choice of the details for the description of each "row" in the forms, if filled in by a specific process or survey, represent a logbook of the actions actually implemented. These could be stored in a database.

Starting from this assumption, we can imagine a tool that "automatically" provides two sets of methodological notes.

The first concerns the usual notes for users. These notes could be supported by an "automatic" extraction of specific information from some of the rows of the form compiled and stored in the DB.

In other words, the description of a specific implementation of the methodological service could be organised as metadata to be inserted in the appropriate section of a documentation note.

Although it is difficult to imagine that complete uniformity can be achieved for all categories of methodological notes and areas of production, as is the case, for example, for European quality reports, it is believed that this would reduce the burden of compiling the notes for process managers.

The second output concerns the automatic drafting of a specific methodological note thought for the producers.

This should contain information belonging to all lines of the process description.

This set of information, if made available to a "new producer", should enable him to reproduce the same results or to implement the same procedure on a new set of input data.

The purpose of this specific methodological note is therefore to guarantee the complete reproducibility of the process considered or to make easier the implementation of the same procedure for a new edition of the same survey by a different team.

4. Final remarks and future developments

4.1 The catalogue

At the very end, several process managers are probably only interested in knowing the availability of IT tools to implement each of the described step of a specific process and to have some hints for their use.

The previous process description does not help very much to this end.

For such reason a new catalogue of methodological methods and tools is under construction.

Such catalogue focusses its attention on the last columns of the forms previously described.

The idea behind this new catalogue is to make the interested researcher able to search for the appropriate method and read the specific descriptions for a "standard" implementation of the related IT tools. In this sense, the catalogue will represent an introductory step into the whole methodological offer available to the production sectors of the Institute.

The design of the catalogue and development of a prototype, since has been carried out by a specialised team, has been heavily relied on the GSIM standard, especially for the identification of objects. Proper descriptions for the most important objects were formulated and their attributes were identified; for example, for the GSIM informative object "functionality" the attributes “name”, “description” and “statistical method” were identified with an accompanying description. As a test, the concepts that were formulated were applied to the SeleMix desktop application.

The prototype has three main functionalities for the correct usage of the tool: a browsing functionality, the access to the methodological services, the management of the catalogue itself (system management).

This architecture has been developed taking into account the heterogeneity of the methods used at Istat: the software tools can be much different in nature that is difficult to abstract and include them in a unique category.

From an IT perspective, the current proposal for the catalogue considers that the methodological services may be offered on a on-cloud application that the users could access to explore the available services and choose the appropriate one. The on-cloud infrastructure would guarantee that the implementation of a service would not be limited by a user’s hardware, since they would be run on an external server. This also would ensure safety standards and scalability. The services would be “wrapped” so that a script could be easily transformed into a statistical service.

4.2 Conclusions and other future developments

As was said in the introduction, the purpose of the activity described in this document is multiple.

In very general terms, it is aimed at improving the efficiency of production processes and the quality of statistics through more effective methodological support. To pursue such a general purpose we need to achieve several partial objectives; in fact, for each step of the production processes we need:

- to share a common language,
- to list the available software tools,
- to indicate the methodologies to be considered as standard,
- to share with production processes the information needed to implement a method or to use a tool,
- to list the potential methodological or technical alternatives available in Institute for each step,

- to supply suggestions, hints, documentation examples and descriptions of previous experiences, and
- to prioritise the research areas and methodological tools development.

To this ends we started from the “atomisation” of the process step having in mind GSIM protocol. Since it is not directly applicable because of several reasons a language closer to spoken methodological language has initially been used.

To achieve a suitable degree of “harmonisation” some “facilitators” worked across specialised teams in charge of specific process ”atomisation”.

It results in the decomposition of the process production phases that should help to pursue the partial and general objectives listed before.

The decomposition of each process with all information collected have been stored in a wiki platform.

The split of the processes has obviously to be considered as an initial step. Once it well accepted by the production sectors the following natural developments should be follow.

The first one is the DB that should contain for each specific implementation of the process phases all information needed to reproduce the same results.

The second one is the full development of the catalogue of the available methods and tools. The third one is the development of the “automatic tool” to fill in the standardised methodological notes.

None of such new developments can be successfully achieved without a fully involvement of the production sectors.

Annex

Table 1 – Excerpt from the template used for process description for some steps of the Integration methodological area

Row number	GSBPM sub-processes	Macrosteps of the process	Summary of the step	Step of the process	Description of the actions
1	5.1 Integrate data	Acquisition of the sources to be integrated	Data reading	Reading of the first sourcew	The data of the first source are loaded along with the names of the existing variables.
2				Reading of the second source	The data of the second source are loaded along with the names of the existing variables.
3		Analysis of the variables	Completeness analysis	Computing of the completeness indicator for a variable	This step is not mandatory and should be carried out only if there are no information on the quality of the variables of the datasets and if no list of the variables to be used for integration has been provided. This indicator computes the frequency by which each variable takes a no missing value. The variables to be preferred in the integration process are the ones for which the value of this indicator is the highest (near 1).
4			Accuracy analysis	Computing of the accuracy indicator for a variable regarding a domain of values	This step is not mandatory and should be carried out only if there are no information on the quality of the variables of the datasets and if no list of the variables to be used for integration has been provided. This indicator computes the frequency by which each variable takes a correct value in the dataset. The variables to be preferred in the integration process are the ones for which the value of this indicator is the highest (near 1).
5			Consistency analysis	Computing of the accuracy indicator for a variable regarding the value of another variable	This step is not mandatory and should be carried out only if there are no information on the quality of the variables of the datasets and if no list of the variables to be used for integration has been provided. This indicator, for a variable, represents the frequency by which the variables takes a correct value if compared to the one of another variable (a correlated variable) on the same row. The variables to be preferred in the integration process are the ones for which the value of this indicator is the highest (near 1).
6			Entropy analysis	Computing of the distribution of the values of a variable in a dataset	This step is not mandatory and should be carried out only if there are no information on the quality of the variables of the datasets and if no list of the variables to be used for integration has been provided. This indicator, for a variable, summarises the frequency distribution of the values of the variable. The variables to be preferred in the integration process are the ones for which the value of this indicator is the highest (near 1).
7			Analysis of the correlation between variables	Computing of the correlation index of two variables in a dataset	This step is not mandatory and should be carried out only if there are no information on the quality of the variables of the datasets and if no list of the variables to be used for integration has been provided. This indicator, for a variable, summarises the correlation of a variable with another variable (defines as correlated variable), that is how much the value of the first variable depends on the value of the correlated variable on the same row. If the value of this indicator is high (near 1) one of the two variables has to be excluded from the set of variables for the matching procedure.

Table 1 cont. – Excerpt from the template used for process description for some steps of the Integration methodological area

GSIM objects							
	Transformable input	**label input**	**Process support input**	**Process method**	**Summary of the transformed output**	**Transformed outputs**	**label output**
1	First archive or source to be integrated	input 1	Metadata for reading (path, record layout, psw etc.)	Data reading (read, read. table, ecc)	Dataset stored in memory	Data from source 1 are available for the elaboration process	output 1
2	Second archive or source to be integrated	input 2	Metadata for reading (path, record layout, psw etc.)	Data reading (read, read. table, ecc)	Dataset stored in memory	Data from source 2 are available for the elaboration process	output 2
3	Loaded datasets, list of variables	output 1 + output 2		Computation of a synthetic indicator	List of variables for matching	Among the variables that are common to the datasets, the ones that have the highest value for this quality indicator and the other indicators will make up the list of the variables for matching, that is the variables that will be used in the next steps of the integration process.	output A1
4	Loaded datasets, list of variables	output 1 + output 2	List of correct values for the variable	Computation of a synthetic indicator	List of variables for matching	Among the variables that are common to the datasets, the ones that have the highest value for this quality indicator and the other indicators will make up the list of the variables for matching, that is the variables that will be used in the next steps of the integration process.	output A1
5	Loaded datasets, list of variables	output 1 + output 2	Name of the correlated variable, list of the correct values for the variable and the correlated variable on the same row	Computation of a synthetic indicator	List of variables for matching	Among the variables that are common to the datasets, the ones that have the highest value for this quality indicator and the other indicators will make up the list of the variables for matching, that is the variables that will be used in the next steps of the integration process.	output A1
6	Loaded datasets, list of variables	output 1 + output 2		Computation of a synthetic indicator	List of variables for matching	Among the variables that are common to the datasets, the ones that have the highest value for this quality indicator and the other indicators will make up the list of the variables for matching, that is the variables that will be used in the next steps of the integration process.	output A1
7	Loaded datasets, list of variables	output 1 + output 2		Computation of a synthetic indicator	List of variables for matching	If the value of the indicator is high (near 1) one between the variable and the correlated variable has to be excluded from the set of variables for matching.	output A1

Table 1 cont. – Excerpt from the template used for process description for some steps of the Integration methodological area

	Methodological tools		
Row number	Statistical methods	Procedures	Process control
1		In R: read.table; read,csv, read csv2 Relais: function dataset	Number of records; variables number, name and type
2		In R: read.table; read,csv, read csv2 Relais: function dataset	Number of records; variables number, name and type
3	Fraction of rows/units with no missing values	Relais: data profiling R: ad hoc function	The best variables are the ones for which the value of the indicator is high (near 1).
4	Fraction of rows/units with a correct value	Relais: data profiling R: ad hoc function	The best variables are the ones for which the value of the indicator is high (near 1).
5	Fraction of rows/units with values correctly linked between the variable and the correlated variable	Relais: data profiling R: ad hoc function	The best variables are the ones for which the value of the indicator is high (near 1).
6	Gini index computed on the frequencies of the values	Relais: data profiling R: ad hoc function	The best variables are the ones for which the value of the indicator is high (near 1).
7	Cramer's V	Relais: data profiling R: ad hoc function	The best variables are the ones for which the correlation with the other variables is low (near 0).

References

Istituto Nazionale di Statistica - Istat. 2016. *Istat's Modernisation Programme*. Roma, Italy: Istat (Accessed 15th February 2023). https://www.istat.it/it/files//2011/04/IstatsModernistionProgramme_EN.pdf.

United Nations Economic Commission for Europe - UNECE. 2019. *Generic Statistical Business Process Model - GSBPM. Version 5.1*. Geneva, Switzerland: UNECE (Accessed 15th February 2023). https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

United Nations Economic Commission for Europe - UNECE. 2021. *Generic Statistical Information Model - GSIM*. Geneva, Switzerland: UNECE (Accessed 15th February 2023). https://statswiki.unece.org/display/gsim/.

Metadata for statistical processes on registers: how to organise facts with GSIM

Mauro Scanu[1], Monica Scannapieco[2], Laura Tosco[1],
Adele M. Bianco[1], Michele K. Riccio[1]

Abstract

This paper illustrates the main concepts for the description of a statistical process based on registers. Most of the concepts are taken from the Generic Statistical Information Model (GSIM, see United Nations Economic Commission for Europe, 2013): even if the standard has been as much as possible preserved, it was necessary to include some additional concepts. These concepts have been inserted in the "meta"-ontological level, one for each fundamental step (layer) of the production process.

Keywords: ontology, GSIM. Micro and macrodata

1. Introduction

Official statistics is making increasing use of registers, which collect and integrate data from both administrative and statistical sources in order to allow the production of a wide range of consistent statistics, while reducing the respondent burden. Having registers as a basis, new and appropriate statistical processes can be put in place.

The objective of this work is to define a layered metadata architecture integrated with the registers' setting that is serving three main purposes:

- Allowing proper statistical analyses on registers' data by providing an explicit and well-defined statistical semantics to microdata present in registers.
- Deriving output statistical datasets from statistical data cubes that, in this case as well, have an explicit and well-defined statistical semantics.

[1] Mauro Scanu (scanu@istat.it); Laura Tosco (tosco@istat.it); Adele M. Bianco (bianco@istat.it); Michele K. Riccio (michele.riccio@istat.it), Italian National Institute of Statistics – Istat.

[2] Monica Scannapieco (m.scannapieco@acn.gov.it), Italian National Cybersecurity Agency.

- Governing the whole production pipeline from registers' data to output datasets by means of a defined and coherent metadata asset.

A first approach has been described in Scannapieco *et al.* (2019), where concept in the *Generic Statistical Information Model* (GSIM, see United Nations Economic Commission for Europe, 2013) have been adopted for assigning a statistical role to metadata. The idea is now to start from the approach in Scannapieco *et al.* (2019) and generalise it to the whole data production process in a registers' setting. This result can be obtained describing a dataset useful for statistical purposes (either a microdata set, as a design matrix, or a table of aggregate data) through a "fact", similarly to the dimensional fact model (see Golfarelli *et al.* 2009): the fact is composed by a measure and intersected semantically by dimensions. In this paper, we detail how a measure should be defined, in a complete and synthetic way, by means of concepts from GSIM (in italics henceforth): more precisely, we distinguish different kinds of measures according to the different kinds of steps available in a statistical process that starts from a register.

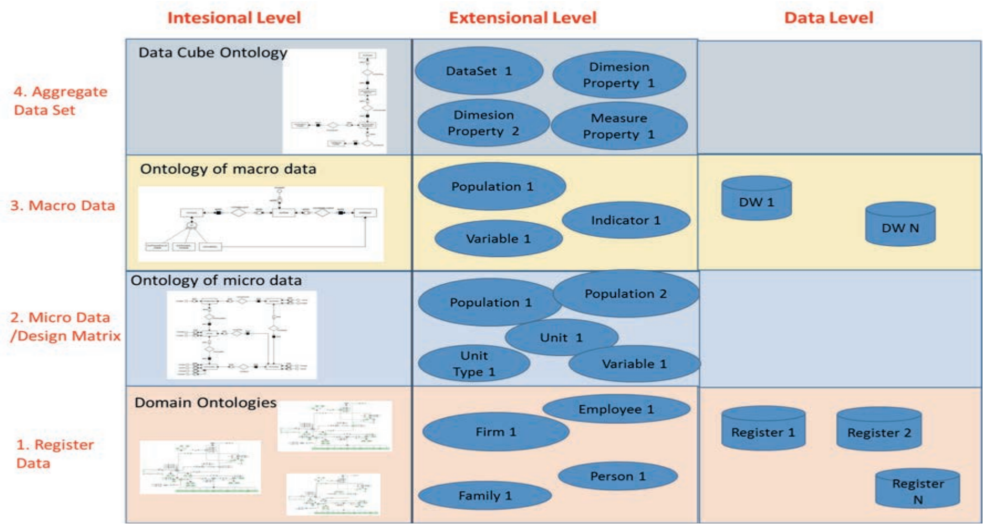
The structure of the paper is as follows. Section 2 described the proposed layered metadata architecture that represents how data in statistical registers can be accessed and published. The subsequent sections give details on each layer of the architecture: Section 3 details the Register layer, Section 4 the Microdata layer, Section 5 the Macrodata layer and finally Section 6 the Aggregate datasets layer. Section 7 was drawn to shade some lights on the nature of statistical data that are managed and Section 8 illustrates some concluding remarks.

2. The different data layers in a register-based production process

What kind of data can be found through a statistical process generated by a register? Figure 2.1 illustrates the answer.

The next sections show in detail the nature of metadata in each layer, and how they should be organised according to the statistical concepts in GSIM.

Figure 2.1 - The four different main layers in a register production process

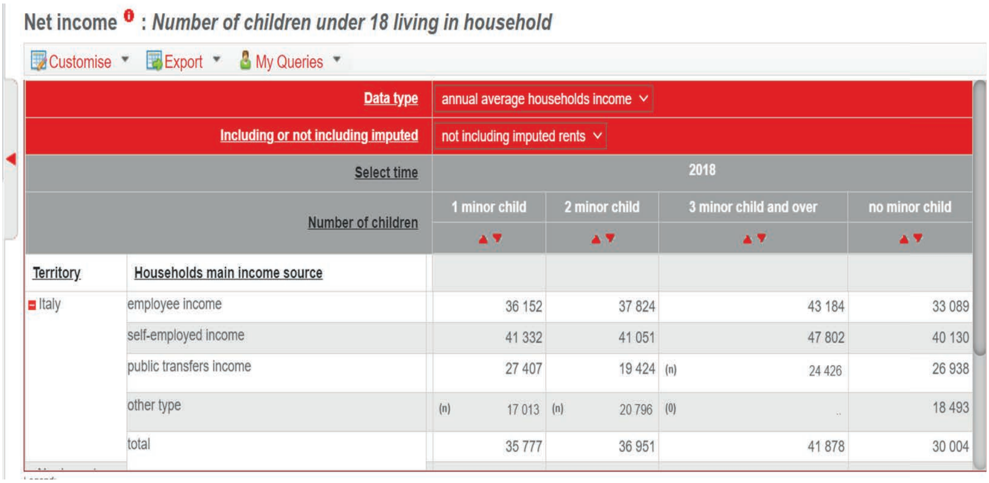


2.1 A case study

In order to show the main features of the four layers, we will follow a case study along the whole process in the next sections. Let us consider the excerpt of two registers under construction in Istat: the income register and the Register of Individuals and Households.

Our objective is modelling all the steps that are necessary in order to publish data like shown in Figure 2.2. This process has the objective to derive these two aggregates: Annual household income without imputed rents and Annual household income with imputed rents. Both the aggregates are disaggregated according to the region of residence in Italy of the household (Territory), the number of children of the household, the household main income source. Finally, the table shows data for the year 2018, anyway other years can be selected. How these output data can be obtained from the two registers, and what consequences can be drawn for the metadata transformation process? Let us start from the initial data available, those organised in the registers themselves.

Figure 2.2 - Table of the annual average household income in Italy per Territory, Number of children and Household main income source in 2018 (and other years), in the two cases in which income includes imputed rents or not



Net income : *Number of children under 18 living in household*

Data type: annual average households income

Including or not including imputed: not including imputed rents

Select time: 2018

	Number of children	1 minor child	2 minor child	3 minor child and over	no minor child
Territory	Households main income source				
Italy	employee income	36 152	37 824	43 184	33 089
	self-employed income	41 332	41 051	47 802	40 130
	public transfers income	27 407	19 424	(n) 24 426	26 938
	other type	(n) 17 013	(n) 20 796	(0) ..	18 493
	total	35 777	36 951	41 878	30 004

Source: Istat (http:\\dati.istat.it)

3. The register layer

The register layer is the initial phase of a register-based data production process. All the relevant data has been collected and appropriately included in the register. From the register, many different datasets can be generated (*i.e.* designed, derived and drawn) according to the appropriate statistical analyses to be produced. Anyway, the register layer is still a too general level in order to be used for just one specific statistical purpose. Data and metadata of a statistical register can be modelled through an ontology. An ontology is a formal, shared and explicit representation of a domain of interest; by simplifying it, an ontology model represents concepts of interest, their attributes and roles existing between concepts.

Our point of view is that the relevant statistical concepts for this layer are those in GSIM in the Concepts section, more precisely *unit type and variable*. Each register is usually characterised by a main *unit type* (*e.g.* enterprises, persons, work positions, education positions), although many others can be found and derived, even if they do not play the role of main unit type. In order to derive easily the datasets on which statistics will be based, a hint could be to design as much as possible the ontology concepts as unit types (especially the main ones). The ontology attributes of each concept play naturally the role of unit type variables. At this level, domain ontologies describe the domain concepts, *i.e.* those concepts that pertain to a specific domain (demography, economy, etc.), independently of the data that are actually available, as well as of the specific statistical interpretation that can be "overlaid" to the data themselves.

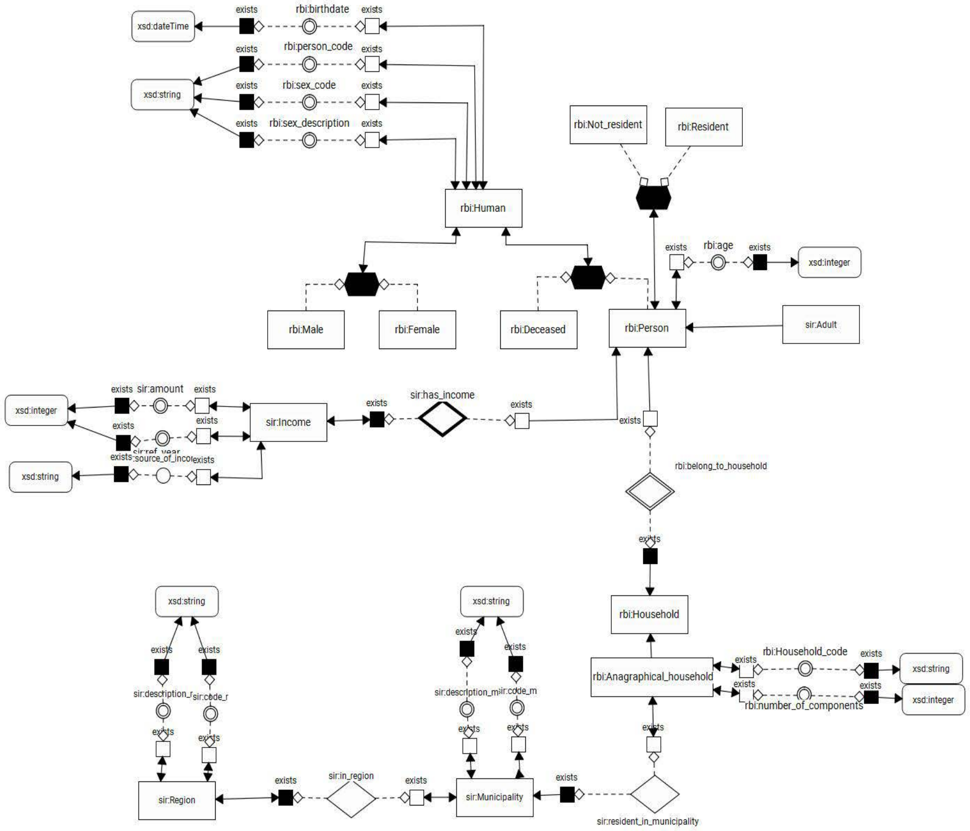
3.1 The case study

In the following we present an excerpt built starting from two registers, the income register and the register of individuals and households, considering just those concepts useful for the case study objective described in Section 2.1. Hence, the ontology considers:

- *Concepts*: Person, Household, Municipality, Income
- *Attributes*:
    - For *Households*: number of children, household identification code.
    - For *Person*: age, birthdate, sex.
    - For *Income*: amount, reference year, source of income.
    - For *Municipality*: code, description.

Why an ontology is better suited for describing the actual content of a register? At this stage, households and persons are not yet organised in populations. A register, if appropriate queries are not applied, contains all the persons collected from all the available data on the topic (administrative archives or other surveys), independently from the fact that these persons are resident or not in a country, if they are dead or alive and so on. Hence, by the ontology, a data analyst has all the relevant concepts that can be studied statistically. The first action that a statistician should perform is a selection (query) from the register concepts and attributes, in order to create the *design matrix*.

Figure 3.1 - excerpt of the ontology on the register layer



Assuming microdata are memorised in a relational database, tables involved in the case study could have the following signatures.

Human (id_person: string, name: string, surname: string, age: integer, adult: boolean, sex: string, birthdate: date, alive: Boolean, resident: Boolean)
Household (id_household: string, anag_household: Boolean, number_of_components: integer)
B_to_h (id_person:string, id_household: string)
Income (id_income: string, id_person: string, amount: integer, ref_year: integer, source: string)
Municipality (code: string, description: string, code_region: string)
Region (code: string, description: string)
H_to_M (id_household_string, id_municipality: string)

To access data through the ontology (OBDA paradigm, Calvanese *et al.* 2009, Poggi *et al.* 2008), we define mapping rules that map data sources to ontology's concepts and roles.

It follows an excerpt of the mapping rules written in GAV (Global-as-view) form (Lenzerini 2002).

```
Person(x,y,z,k,l,s,b,r) <-- select id, name,surname, age,adult,sex, birthdate,resident from Human
                            where alive=TRUE
Household (x,y,z) <-- select * from Household
belong_to_household (x,y) <-- select * from B_to_h
Income (x,y,z,k) <-- select id, amount, ref_year, source from Income
has_income (x,y) <-- select id_person, id_income from Income
Municipality (x,y) <-- select code, description from Municipality
resident_in_municipality (x,y) <-- select id_household, id_municipality from H_to_M
in_region (x,y) <-- select code, code_region from Muncipality
Region (x,y) <-- select code, description from Region
```

3.2 How design matrices can be built from register layer ontologies

Statisticians are used to deal with very simplified data representations at the unit level: on the one hand, they refer to units in a population, on the other on variables observed on the population units. The result is a rectangular matrix of data, named *Design Matrix*, with as many rows as the units in the population and as many columns as the variables observed on the units. The generic element in the *i*-th row and *j*-th column is the value/item of variable *j* observed on unit *i*. Design matrices can be designed from a register layer by picking one concept in the ontology as a *unit type*, and deriving *unit* and *population* by making explicit assumptions on time, place and possibly other concepts: *e.g.* from the unit type "person" it is possible to derive the population of individuals *in Italy - on the first of January 2020 – who are residents* (so that from the generic list of persons, the set of units of interest for the statistics are restricted to the place Italy, the time first of January 2020 and, whenever necessary, the additional features as the fact to be resident in a country). As a matter of fact, it is possible to derive other unit types by appropriately querying concepts and attributes.

As far as the variables are concerned, the most straightforward set of variables are the attributes of the chosen unit type. This can be a first ready-to-use design matrix. Anyway, the ontology offers many other characteristics of the unit of the population of interest that can give rise to variables, enhancing the multivariate observed on the population units. These additional variables can be derived exploiting the relationship between the unit type that characterises the unit/population of interest and the other *concepts* in the ontology.

- One-to-one relationship between *unit types* (*e.g.* person and resident in Italy): in this case, the attributes of one *unit type* can be *variables* of a *unit* defined from the other *unit type*. For instance, gender (an attribute of person) can be a variable for the residents in Italy on the 1st of January 2020, while the residential address (an attribute of the residents in Italy) can be a variable for the corresponding person.
- One-to-many relationship between *unit types* (*e.g.* household and person): the attributes of the "larger" *unit* can play the role of a *variable* for the "included" *units*. For instance, a person that belongs to a household inherits all the household attributes (household telephone number, address, square metres of the house, etc.). It is important to add a statistical alert: household variables for the *units* belonging to the same household are not independent in the probabilistic sense: for these variables, statistical methods known as "multi-level models" can be appropriate.
- Many-to-one relationship between unit types (*e.g.* person and household): in this case, it is necessary to make a preliminary step of aggregation up to the "larger" *unit*. For instance, at the household level it is possible to derive:
    - through enumeration: the number of individual components;
    - through enumeration by means of other (categorical) attributes at the person level: the number of males (through gender), the number of those with a university degree (through educational level), the number of people in the household aged 60 or older (through- age);
    - by means of other statistical analyses on attributes: the main income component, the average height, the median educational level.

Hence, a design matrix for a unit type can potentially take advantage of the information coming from the whole set of concepts in an ontology and their attributes. For this reason, the register layer can be thought as the "mother of all the possible traditional design matrices" derivable from the register, each design matrix devoted to specific statistical purposes. Once understood as a design matrix can be built from the ontology concepts, a statistician can start working in a well known environment, where the objective is to build aggregate data from observations in a microdata set.

4. The microdata set (or design matrix) layer

The microdata set level can be defined starting from the register layer as follows:

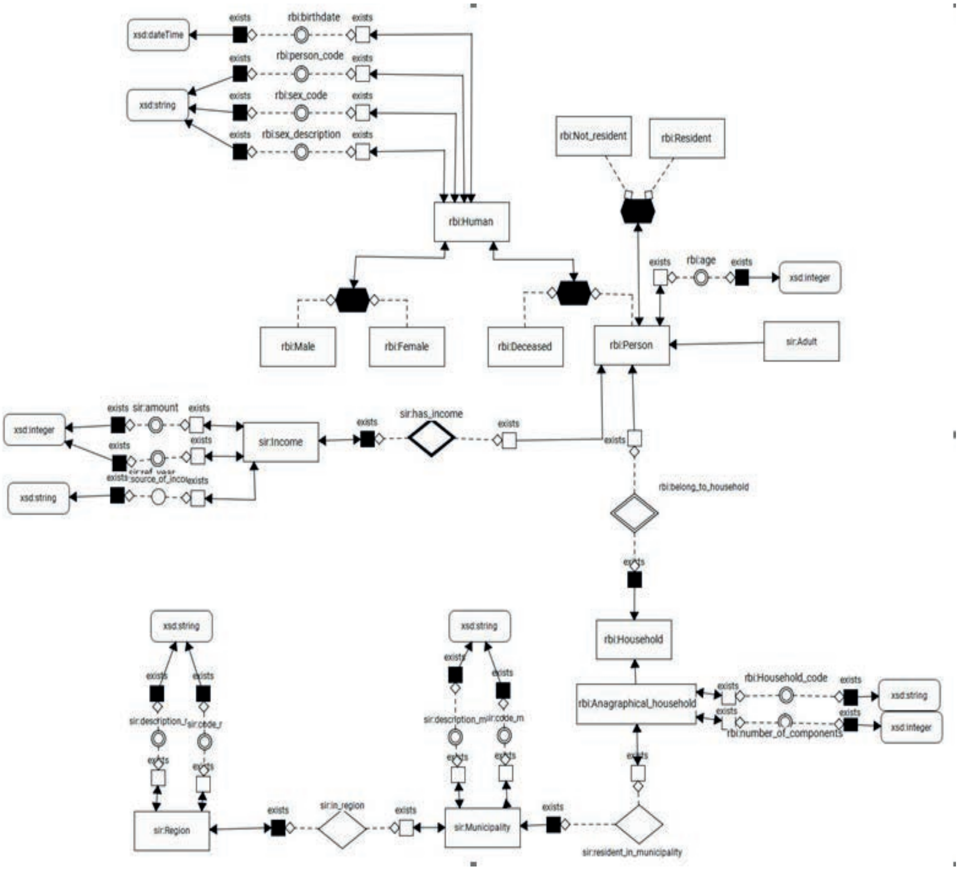
- *Universe*: *i.e.* a query between a *unit type* concept and its possible linked or derived *variables*;
- *Population*: *i.e.* a *universe* with a specific time and space location of the units in the *universe*;
- *Represented variables*: *i.e*. the characteristics of interest observed on each unit of the *population/universe*.

As already explained in Section 2.1.1, the design matrix is a matrix where each row represents and is associated to a unit in the population, and each column to a variable observed on each unit, so that the datum in a cell of the design matrix corresponds to the category/value of the variable represented in that column observed on the specific unit in that row. The list of design matrices for all the spaces (geographical areas) and times defines all the data in the universe of interest. In other words, a design matrix can be described as a data structure, constrained to include microdata, *i.e*. the input data in a statistical process. Note that the set of rows corresponds to the whole set of units in a population. If this is not the case, the units in the list should be a representative portion of the population (for the sake of simplicity, we will restrict to completely observed populations, without harming the metadata representation of data in the different layers; see Section 7 for this issue). Usually, if we include in the design matrix the territorial/geographical variable as one of the variables of interest for the units in the population, we end up with multivariate cross-sectional analysis of a population in different time points. If we include the time dimension in the design matrix, we get a dataset on a cohort of statistical units along time, useful for longitudinal data analyses.

At this level, a meta-ontology at micro level has been introduced with the purpose of providing the statistical semantics above described (see Figure 4.1). The most important aspects of this meta-ontology are

- The definition of *universe* and *population* concepts in terms of *unit type* and time and geographical *attributes*;
- The inclusion of concepts relative to the multivariate variables jointly observed on each unit.

Figure 4.1 – the microdata set layer ontology



4.1 The case study

The design matrix for the case study introduced in Section 2.1 consists of the following characteristics:

- Unit type: Household.
- Population: the whole set of resident households (Universe) in Italy (Geographical Area) in the first of January 2018 (Time).
- Unit: each resident household in Italy in 2018.
- (Joint) Represented Variables: Amount of income, Amount of income with imputed rents, number of children, description of municipality (and consequently province and region), income source.

Data in Figure 2.2 is only a slice of the hypercube in the Istat corporate data warehouse, where other populations consists of the same universe and geographical area, in different times (years other than 2018): these populations are different year after year (for deaths, newborns, and population movements in and out a country). If on the contrary the focus is longitudinal, once selected a cohort in a specific time, this set of units is followed year after year. Also in this case, deaths and drop outs may occur, anyway, unless planned, no other units are added on the cohort.

The Unit Type concept is instantiated with the "name" of the concepts of the domain ontologies (*i.e.* Household, Person, Municipality). Represented Variables concept is instantiated with the name of the attributes of the concepts of the domain ontology; *i.e.* amount of income is an instance of the concept Represented Variable and is, in the domain ontology, the name of an attribute of the Income concept. This is the first way of "connection" between metadata layer and the underlying microdata layer.

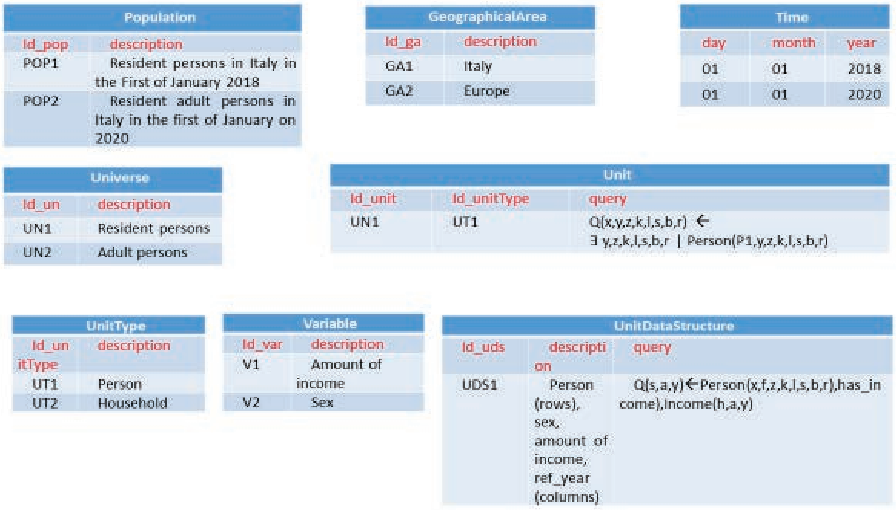
Let us suppose metadata are memorised in a relational database and relevant tables for the case study have the following signatures.

```
Population(id_pop:string, description:string)
GeographicalArea(id_ga:string, description:string)
Time(day:integer, mounth:integer, year:integer)
Universe(id_un:string, description:string)
Unit(id_unit:string, id_unitType:string, query:string)
UnitType(id_unitType:string, description:string)
UnitDataStructure(id_uds:string, description:string, query:string)
Variable(id_var:string, description:string)
```

The “query” attribute both of the Unit and Unit Data Structure concept, is very important: it performs the second way of “connection” between metadata layer and the underlying microdata layer. This attribute contains the SPARQL queries over the domain ontologies (first level of our architecture) that allows the microdata extraction.

Even at this level, to access metadata through the ontology we have to write the mapping rules of which an excerpt follows.

Figure 4.2 – example of table instantiation related to the case study ontology



Population	
Id_pop	description
POP1	Resident persons in Italy in the First of January 2018
POP2	Resident adult persons in Italy in the first of January on 2020

GeographicalArea	
Id_ga	description
GA1	Italy
GA2	Europe

Time		
day	month	year
01	01	2018
01	01	2020

Universe	
Id_un	description
UN1	Resident persons
UN2	Adult persons

Unit		
Id_unit	Id_unitType	query
UN1	UT1	Q(x,y,z,k,l,s,b,r) ← ∃ y,z,k,l,s,b,r | Person(P1,y,z,k,l,s,b,r)

UnitType	
Id_unitType	description
UT1	Person
UT2	Household

Variable	
Id_var	description
V1	Amount of income
V2	Sex

UnitDataStructure		
Id_uds	description	query
UDS1	Person (rows), sex, amount of income, ref_year (columns)	Q(s,a,y)←Person(x,f,z,k,l,s,b,r),has_income),Income(h,a,y)

```
Population(x,y) <–– select * from Population
GeographicalArea(x,y) <–– select * from GeographicalArea
Universe(x.y) <–– select * from Universe
Unit(x,y) <–– select id_unit, query from Unit
isOfTypeU-UT(x,y) <–– select id_unit, id_unitType from Unit
UnitType(x,y) <–– select * from UnitType
Variable(x,y) <–– select * from Variable
UnitDataStructure(x,y,z) <–– select * from UnitDataStructure
```

Briefly, step between the register and the design matrix layers is structured as follows:

- Select the unit types.
- Select the population for the unit types: a specific time and a specific geographical area in which the units that correspond to that unit type exist (in the case study, 1-1-2018 and Italy) and any other characteristic that is necessary for the definition of the population (in the case study, the households should be resident in the cho-sen geographical area).
- Select the useful variables, and derive new ones, following the rules of Section 3.2:
    - number of children,
    - household residence,
    - household annual income per income source (derived variable obtaines summing each person household annual income per income source),
    - household main income source (derived looking at the income source for which there is the maximum among the different household annual income per income source),
    - household annual income (sum of the household annual income per income source),
    - household annual income with imputed rents (the household annual income is increased of imputed rents for those household that do not belong the residence house).

5. The macrodata (indicator) layer

The design matrix is the finest and most complete source of statistical information on the joint set of variables on the selected population. Anyway, information on the different specific units that compose the population is usually redundant for statistical purposes. The most compact way to represent the whole statistical information in the design matrix is the empirical cumulative distribution function. it is obtained by:

- computing the multivariate contingency table for all the categorical variables in the design matrix;
- given each combination of the categorical variables, the joint cumulative distribution of the numerical variables completes the computation.

This source of information is seldom represented as a result in official statistics, while specific summaries of the empirical cumulative distribution function are actually disseminated: as already said these numbers are usually named “aggregated values”, *i.e.* computed with respect to sets of units. The aggregated values represent how the variables distribute over the population of interest. The way to derive these aggregated values depends on the nature of the variables themselves. Usually, the output of a statistical analysis from a National Statistical Office, consists of the following three frameworks:

a. *The multivariate variable under study consists only of categorical variables*. If variables are only categorical, the corresponding empirical distribution function can be represented just by contingency tables (or their direct transformations as percentage distributions). Hence, enumeration or percentages are the direct representation in aggregate terms of a distribution of a multivariate categorical variable in a population. This representation does not loose statistical information with respect to the microdata layer of Section 2.2 (only the useless unit identifiers are lost, without harming information).

b. *The multivariate variable under study consists only of categorical variables, but the main interest is just on one or a few of them.* If, among the categorical values, the focus is only on one of them, while the other variables categorise the different subpopulations on which computations are built, the result is usually the percentage distribution of the focus variable given the other categorical variables (*e.g.* percentage of smokers in the subpopulations given by gender, age class, educational level). Hence, only the distribution of the focus variable given the other categorical (conditional) variables is actually considered, while the distribution of the conditional variables is not represented (and this represents a loss of information with respect to the joint univariate observation available in point 2).

c. *The multivariate variable includes one numerical variable.* If among the variables there is at least one numerical variable, the corresponding empirical distribution function conditional to the different values that the joint categorical variables can assume is difficult to represent exhaustively. Hence, these distributions are usually represented by specific characteristic distribution values. Restricting the attention to a univariate numerical variable, as usually happens in official statistics, these values can be, mentioning the most used values in official statistics:

- the "position parameters" of a distribution (the mean, the median, the percentiles, the minimum, the maximum);
- the "dispersion parameters" of a distribution (variance, coefficient of variation, interquartile range);
- the homogeneity parameters (*e.g.* the Gini index).

Anyway, the same way of reasoning applies for multivariate numeric variables, even if the possible outputs are out of the scope of National Statistical Institutes: *e.g.* correlation coefficients, linear regression parameters, etc. In this case, with respect to the design matrix, the choice of the aggregate value corresponds to the loss of the distribution of the categorical variables as well as of the "form" of the distribution of the numerical variable(s) given the categorical ones with the exception of the characteristic values that have been chosen.

5.1 The case study

The case study in Section 2.1 represents two aggregate values: the annual average household income with and without imputed rents. Let us restrict to the one without imputed rents. The aggregated value is obtained through the composition of the following instantiation of concepts:

- The whole set of resident households: instance of Universe;
- Amount of income without imputed rents: retrieved by the SparQL query inside Joint_variables concept;
- Average: instance of Rule.

These three instances of concepts give the main information that is going to be represented in a table of aggregated data. Other metadata available in the case study of Section 2.1 specifies better the meaning of each datum in the cells of the table, anyway do not give further information on the meaning of the aggregated value represented in the table. They will be further studied in Section 6. From Figure 5.1, representing the ontology of the macrodata layer, let us instantiate the necessary concepts according to what described in the previous example. For the sake of space, in the following there is the signature of relational tables related to the main concepts in this case study.

Indicator	
Id_i	Description
I1	“Average household income without imputed rents”

Joint_variables	
Id_jv	Query
JV1	Q(k)ß $ x, y, z, k | UnitDataStructure(x, y, z) , hasDS-MC (x,k) , MeasureComponent (k, ”income without imputed rents”)

Process_Method			
Id_pm	applied_on	hasPM-R	Has_variable_set
PM1	U1	R1	JV1

Universe	
Id_u	Description
U1	“resident households”

Rule	
Id_r	Rule_name
R1	“Average”

PM_defines_ind	
Id_pm	Id_i
PM1	I1

With mapping rules:

```
Indicator (x,y) <–– select * from Indicator
Joint_variables (x,y) <–– select * from Joint_variables
Process_Method(p) <–– select * from Process_Method
Universe(x.y) <–– select * from Universe
Rule(x,y) <–– select * from rule
defines(x,y) <–– select * from PM_defines_ind
applied_on(x,y) <–– select Id_pm, applied_on from Process_method
has_PM-R_on(x,y) <–– select Id_pm, hasPM_R from Process_method
has_variable_set(x,y) <–– select Id_pm, has_variable_set from Process_method
```

5.2 Example relative to the other two aggregate value frameworks

The characteristics features of a multivariate distribution can be defined by appropriately selecting through a query the following concepts: *Universe, Quantitative represented variable, Process method*. The case study gives an example of what happens in the framework *The multivariate variable includes one numerical variable*. Let's see how the other two aggregate value frameworks described in Section 5 adapt to this query structure.

- *The multivariate variable under study consists only of categorical variables.* This case is very similar to the one shown in the case study, anyway there is a specific caution to consider for the *Quantitative Represented Variable.* Furthermore, the actual focus of the aggregate is in the joint distribution of the whole set of categorical variables, that should be clearly declared. Let us consider as an example the aggregate: Resident population in Italy by gender, age and marital status.
    - *Universe*. Resident population (persons who have a residence in a specific geographical area).
    - *Quantitative Represented variable*. When the analysis is only on categorical variables, the quantitative variable is usually omitted and corresponds to the "Counting measure" that enumerates the units in the population with the specific characteristics of the Categorical Represented Variables. This variable remains unchanged for all the tables consisting of contingency tables.
    - *Process method*. The objective is to get the total enumeration of the units in the population with specific characteristics; hence the *Process method* is Total. If the method was the percentage, each enumeration has to be divided by the number of units of the corresponding population.
    - *Categorical Represented Variables*: gender, age and marital status.

- *The multivariate variable under study consists only of categorical variable, but the main interest is just on one or a few of them.* An example is in Figure 5.2: what is the aggregate relative to, for instance, the first number (19.8) in the upper left part of the table? According to the available metadata, that number represents the percentage of variable Smoking habit. What is the role of the other variables and how can we model it?

- *Universe*. Resident population aged 14 years and over.
- *Quantitative Represented variable*. Counting measure.
- *Process method*. Conditional percentage.
- *Main Categorical Represented Variables*: smoking habit (with associated code list consisting of the categories: smokers, former smokers, ever smokers persons).

Figure 5.1 - The macrodata layer ontology

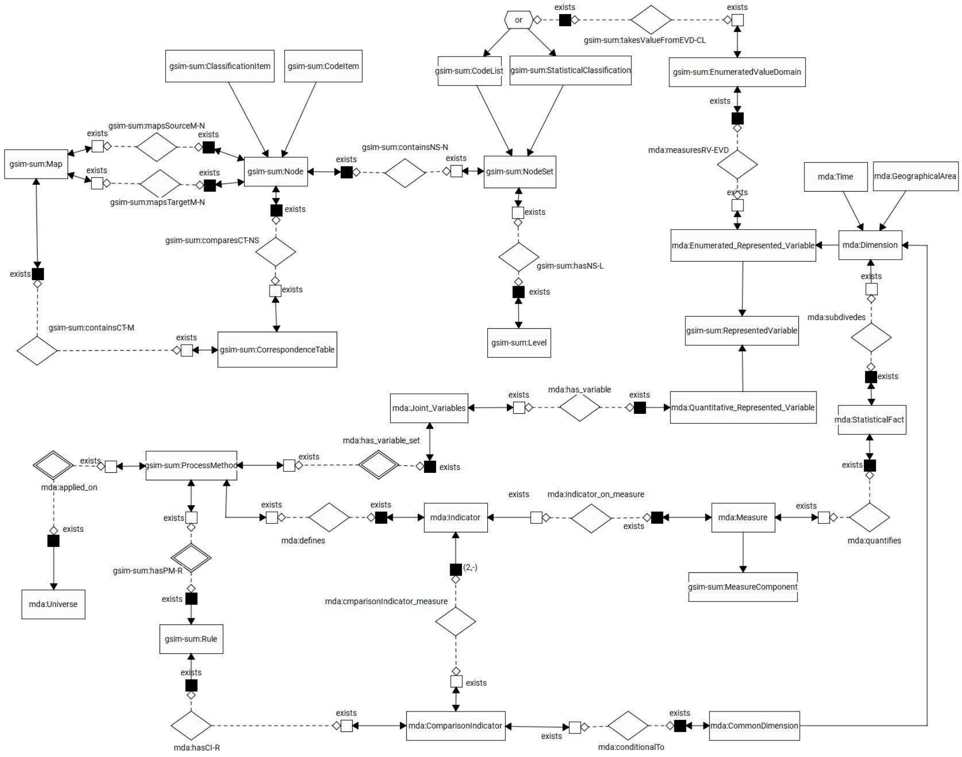
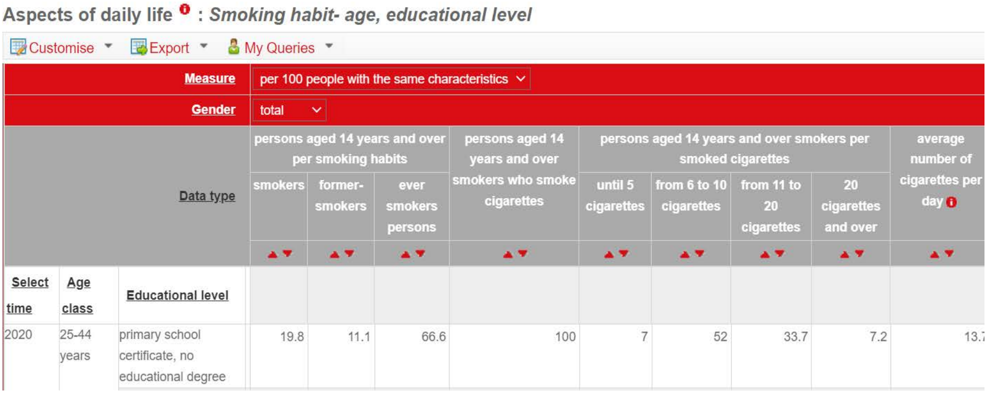


Figure 5.2 - Example of macrodata with many aggregates, mostly where one categorical variable is the focus



Aspects of daily life : *Smoking habit- age, educational level*

Measure: per 100 people with the same characteristics

Gender: total

		Data type	persons aged 14 years and over per smoking habits			persons aged 14 years and over smokers who smoke cigarettes	persons aged 14 years and over smokers per smoked cigarettes				average number of cigarettes per day
			smokers	former-smokers	ever smokers persons		until 5 cigarettes	from 6 to 10 cigarettes	from 11 to 20 cigarettes	20 cigarettes and over	
Select time	Age class	Educational level									
2020	25-44 years	primary school certificate, no educational degree	19.8	11.1	66.6	100	7	52	33.7	7.2	13.7

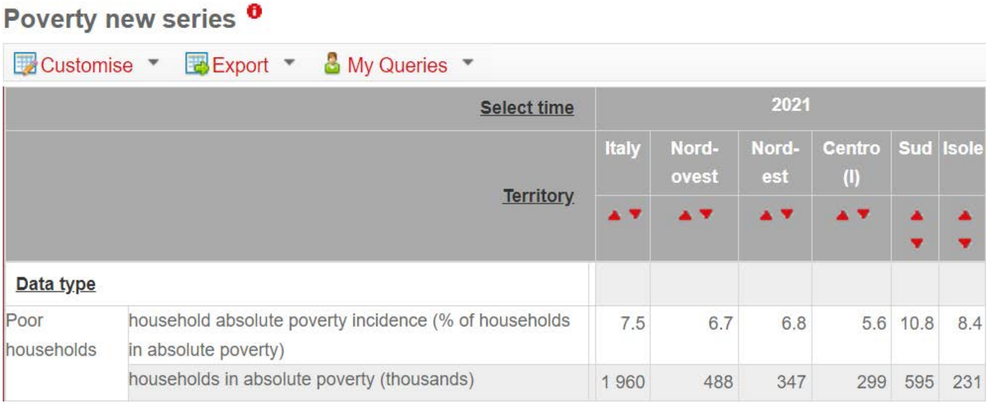
Source: Istat (http:\\dati.istat.it)

5.3 Other statistical outputs: Comparison indicators

Usually, a statistical process does not end with the representation of the distribution of variables on one population, whose three frameworks in Section 5 are some of the possible outputs. Other analyses are performed, mainly in terms of comparisons. Comparisons are among already computed aggregate values: differences, ratios, densities, index numbers, percentage variations are typically computed by comparing already aggregated data by simple algebraic computation (differences by means of subtractions, ratio and densities through a division) or by a combination of ratios and differences (percentage variations) or by ratios that need the computation of intermediate values (*e.g.* the product of prices and quantities in different time points as for index numbers). Examples are comparisons that pinpoint parts of a population (*e.g.* the absolute poverty incidence), or the comparison of aggregates computed on the same population but different variables (*e.g.* propensity to consume), or different population and variables (as the ratio between the kilograms of waste produced in a municipality divided by the number of inhabitants in the municipality), or same variable and population at different times (as for index numbers).

This level also needs the definition of a meta-ontology, *i.e.* a macrodata ontology, taking into account the need of representing aggregated values and the characteristics of the variables contributing to the aggregation. See Figure 5.1, and note the specification of the concept *measure*, that correspond to the core of aggregate value, and its relationship with the microdata layer concepts (*represented variable* with a numeric *value domain, population, etc.*).

Figure 5.3 - Example of a statistical output consisting of a comparison indicator



Poverty new series

	Select time	2021					
	Territory	Italy	Nord-ovest	Nord-est	Centro (I)	Sud	Isole
Data type							
Poor households	household absolute poverty incidence (% of households in absolute poverty)	7.5	6.7	6.8	5.6	10.8	8.4
	households in absolute poverty (thousands)	1 960	488	347	299	595	231

Source: Istat (http:\\dati.istat.it)

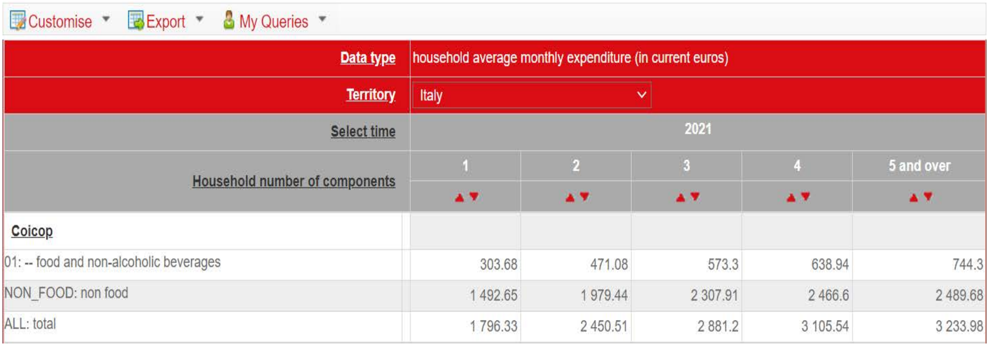
5.4 Example

Figure 5.3 shows the case of a Household absolute poverty incidence. According to information on the Istat data warehouse (http://dati.istat.it): “The Istat estimate of the absolute poverty defines as poor a household with a consumption expenditure lower or equal to the monetary value of a basket of goods and services considered as essential to avoid severe forms of social exclusion. The monetary value of the basket of absolute poverty is reviewed every year in the light of trend in prices and compared to the levels of spending on household consumption. [...] The proportion of poors (incidence) [...] is the ratio between the number of households (individuals) in poverty and the number of resident households (individuals)”.

Hence, each output in the first line of Figure 5.3 can be obtained dividing each number in the second line (got through the already available design matrix once “the monetary value of a basket of goods and services considered as essential to avoid severe forms of social exclusion”) with the corresponding number of households in Italy and each repartition, for instance from the Census.

How the second line in Figure 5.3 has been obtained? We need to consider the distribution of expenditures per resident household in Italy in 2021 for all the goods and services in the basket. Microdata exist and produce, for instance, data in Figure 5.4.

Figure 5.4 - A table of data representing the household average monthly expenditure in Italy (in current Euros)



Data type	household average monthly expenditure (in current euros)				
Territory	Italy				
Select time	2021				
Household number of components	1	2	3	4	5 and over
	▲▼	▲▼	▲▼	▲▼	▲▼
Coicop					
01: -- food and non-alcoholic beverages	303.68	471.08	573.3	638.94	744.3
NON_FOOD: non food	1 492.65	1 979.44	2 307.91	2 466.6	2 489.68
ALL: total	1 796.33	2 450.51	2 881.2	3 105.54	3 233.98

Source: Istat (http:\\dati.istat.it)

Data in Figure 5.4 represent only a specific characteristic of the distributions of expenditures on the resident Italian households in 2021: the average. It is important to underline a characteristic of this table, according to what already said in the sections before. Figure 5.4 includes as many aggregate values as the categories in the dimension “Coicop”: for the sake of convenience, we reported only the Figures for food and non-food items, as well as the total, anyway the Coicop code list has got many other items useful in order to fill in the total expenditure of each household in the basket. Coicop is not, strictly speaking, a classification because it does not classify units of a population (in this case the households). Coicop lists the numerical variables observed on the households in microdata and for which aggregate values (the averages over the households) are given. Hence, in order to get the numbers in the second line of Figure 5.3 it is necessary to start from the register, have the first query on the population and the variables of interest (that will include all the numerical variables relative to expenditures on goods and services in the basket), consider the derived variable given by the sum of the selected expenditures for each household, and finally derive an indicator variable that compares the threshold under which a household can be considered as absolutely poor or not. Aggregation (total) with respect to this final derived variable will give the results shown in Figure 5.3.

6. The aggregate dataset layer

The three aggregate value frameworks representing (part of) a statistical distribution described in Section 5 apply for all the values of the categorical variables available in the joint multivariate distribution (as represented in the design matrix of Section 3).

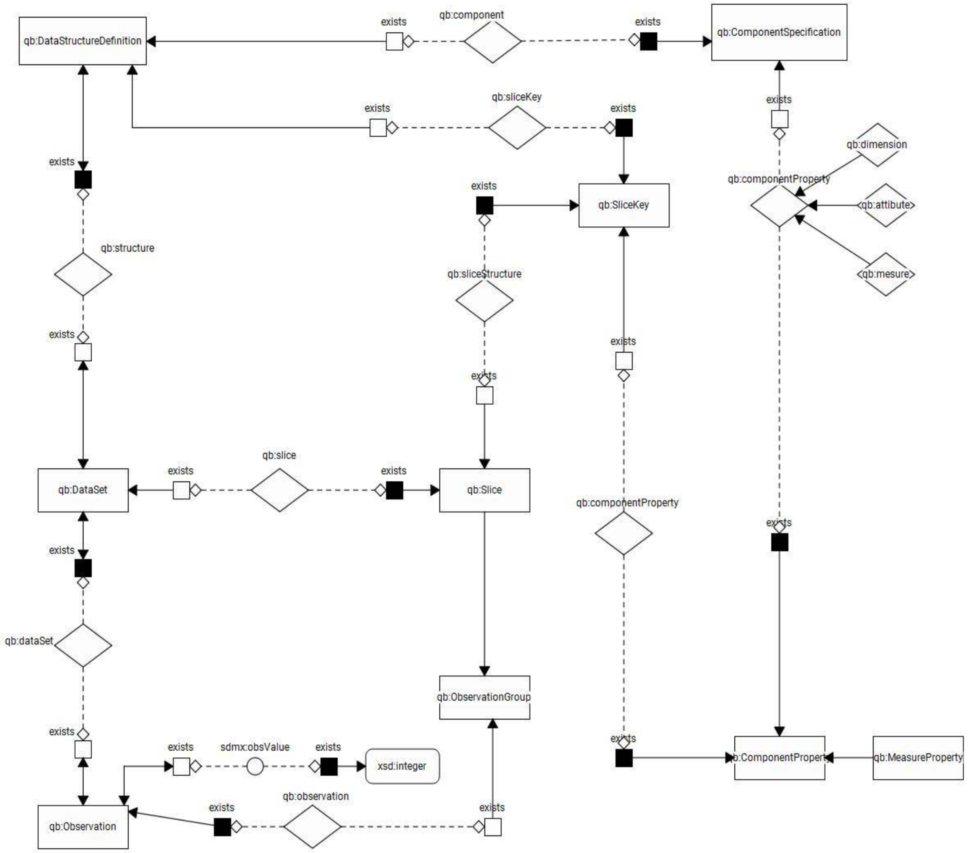
Considering all the values assumed jointly by the categorical variables, it is possible to represent the whole dataset of aggregate values (each one determining the *datum in a data point of a unit data structure*).

a. In case there are only categorical variables, the table represents the enumeration of the units of a population according to the joint set of categorical variables of interest: hence, the dimensions of the dataset should contain all the categorical variables.

b. In case the focus is only on one categorical variable, the dataset will usually show the percentage of an item in a classification (*e.g.* smokers), or the complete distribution of the focus variable for all the categories in the classification (*e.g.* smoker/non-smoker), for the combination of all the other categorical variables available. The conditional variables should be included as dimensions in the dataset, representing the subsets of populations on which the represented percentage should refer to. The focus variable can be either a dimension (especially when the whole conditional distribution is represented) or not (when the percentage of only one category is represented).

c. In case of a characteristic value of a numeric variable given the categorical ones, the dataset should include the categorical variables as dimensions, so that the dataset represents the characteristic value for each subpopulation given by the combination of specific values of the classification items.

Figure 6.1 – The aggregate dataset ontology



The resulting datasets are nothing but queries over the whole cubes defined by the macrodata ontology. However, these cubes deserve a specific representation for practical reasons: National Statistical Institutes publish time datasets as their principal outputs. The specific design choices related to the output, need to be represented. Here, already defined meta-ontologies, like *e.g.* the RDF Data Cube Vocabulary ontology (see World Wide Web Consortium 2014) can be reused (see Figure 5.1).

We can formally derive the components of a data structure by means of the already introduced concepts from GSIM. The data structure is described by these concepts:

- *Measure*: a *Measure* consists of the aggregate/indicator as defined in Section 5. A dataset can report more than one measure: in this case it is necessary to introduce a specific dimension representing what is measured in the table, that includes all the indicators to show in the dataset. If the dataset show data on the aggregate framework b), the corresponding *Measure* (*i.e.* the percentage of the variable focus of the analysis) is represented by the code list of categories of the focus variable.
- *Dimension*: What is measured in the dataset is "decomposed" by different kind of variables:
  - *Categorical Represented Variables*: just those that are not the focus of aggregate framework b). The joint set of values of these variables decompose the *Population* in distinct sets over which the aggregate value should be computed.
  - *Time variable*: this variable lists all the populations that can be derived from the Universe defined in the aggregate value (*Measure*) according to time.
  - *Geographical Area Variable*: this variable lists what is the main geographical area of the population that characterises the *Universe* (*e.g.* Italy) and all the subpopulations over which the aggregate is computed (*e.g.* ripartitions, regions, provinces, municipalities, etc.). Hence, it has the same role as any *Categorical Represented Variable* (where Italy corresponds to Total), with the major characteristic that this variable is ubiquitous in the datasets of any disseminated statistics. Note that this variable is "time dependent", *i.e.* identifies the partition of the population according to the time instant defined in the Time variable (*e.g.* the municipality of Sappada should be included in the region Veneto up to 2017, and Friuli-Venezia Giulia from 2017 onwards).
- *Attributes*: any other characteristic that is not part of the definition of the meaning of the dataset cell, that helps the understanding of the cell content (*e.g.* if the represented value is provisionary or definitive).

The combination *Measure-Dimension* is the definition of the content of each cell in the dataset, usually takes the name of *Fact*, as defined in Golfarelli *et al.* (2009), in IT studies. So, a statistical *Fact* should always consist of an aggregate value of a distribution on a population or an indicator, per the

Categorical Represented Variables, *Time and Geographical Area* variables, nothing more, nothing less.

6.1 The case study

The table represented in the case study of Section 2.1 represents data relative to two aggregates. One has already been described the aggregate described in Section 5.1, *i.e.*:

- *Universe*: the whole set of resident households;
- *Quantitative represented variable* (if present): Annual income without imputed rents;
- *Process method*: average;

while the other aggregate is similar but with Quantitative Represented Variable given by Annual income with imputed rents. Hence, the table represents two aggregates given by the combination of (in Figure 2.2 this is given by the combination of the dimension "Data type" and "Including or not including imputed rents". These two aggregates are then disaggregated in a number of Figures with specific meaning of each datum in the cells of the table. These further information is given by Time and Geographical area (that specify the Universe into specific populations) as well as the other categorical variables that partition the population in distinct sets of units. For instance, it is possible to write that the aggregated value that is represented in the table in Figure 2.2 is the average (Process) resident household (Universe) annual income without imputed rents (Quantitative represented variable) PER or GIVEN, *i.e.* having fixed the value of, and selecting only those households with a combination of:

- *Time*: *e.g.* 2018 or any other available year that better specifies the population on which the aggregated value is computed;
- *Geographical area*: Italy and all areas in the code list of Residence included in Italy;
- *Categorical Represented Variables*: number of children, main income source.

One of the *Facts* in the dataset is: Average resident household annual income without imputed rents in Italy and its main disaggregations, per

Year, Number of children and Main income source (for the other one, change "without" in "with").

For the examples introduced in Section 5.1 for describing what happens in the different frameworks, the additional concepts to be used in order to represent the aggregate in its table are:

- *The multivariate variable under study consists only of categorical variables*. *I.e.* for the example: Resident population in Italy by gender, age and marital status.
  - *Time*: currently Istat corporate data warehouse disseminates data that refers to the 1st of January 2022, and the years before;
  - *Geographical area*: Italy (and all areas in the code list of Residence included in Italy).
- *The multivariate variable under study consists only of categorical variable, but the main interest is just on one or a few of them*. By means of the example in Figure 5.2 we already know that the aggregate relative to the first Figure (19.8) in the upper left part of the table is the percentage of variable Smoking habit observed in a subset of the Universe of the resident population aged 14 years and over. Anyway its correct and complete interpretation still needs information on how these subsets are created. These are the relevant concepts to include in the description of the table of data additionally to the already given aggregate.
  - *Time*: 2020, and the years before;
  - *Geographical area*: Italy;
  - *Other (conditional) Categorical Represented Variables*: age class, educational level, gender.

Hence, the first Figure in the upper left part of the table represents the percentage of smokers (19.8%) in the part of the population (residents aged 14 years and over in Italy in 2020) determined by the other categorical variables, *i.e.* age class (25-44 years), educational level (primary school certificate, no school degree) and whatever gender (total).

7. Counts or estimates?

So far, we have studied the structural metadata associated to data in the different layers of a data production process, from the register level up to the dissemination of aggregates and indicators in a dissemination table, without declaring if the actual computation of the aggregate should follow a computation on a complete set of observations or on just a sample. This is because structural metadata do not declare the features of the statistical process in such detail, focussing on the objective that the corresponding data *should mean*. Data in Figure 2.2 are actually estimates from a sample of observations: this characteristic is declared elsewhere. On the contrary, the disseminated table only declares that the Figures in the table are relative to the Average annual household income.

Is it possible to include information on the computation process (if by counting or estimation) in the metadata? As a matter of fact, the process we have developed allow to dedicate a specific step to this issue.

The crucial point is in the construction of the design matrix. If the design matrix is complete, *i.e.* has as many rows as the corresponding population of interest, we are in the case of a traditional census case: hence computation (totals, averages, medians, etc) can be computed directly on the whole set of available data. If, on the contrary, the design matrix, *i.e*. the result of the query in section 3.2 does not contain all the records in the population, it is necessary to follow these steps:

1. be sure the available data is representative of the population of interest. In case it is, verify if there is a need to add an additional column to the design matrix, *i.e.* the one relative to a survey weight in order to correctly weight the observation in the computation of the aggregates;
2. use an appropriate estimator (for instance, one that allows for good features of the mean square error in some important cases).

Hence, by appropriately checking the step of the design matrix creation, this aspect can be included among the metadata to attach to the data. This approach will be further studied elsewhere.

8. Conclusions

In our opinion, the main steps to consider in a statistical process based on registers are those represented in Sections 3-6. As a matter of fact, the process seems to be endless, aggregated datasets can be actually (even partially) used as micro datasets at different unit types (for instance, we can use municipalities as statistical *units* and the aggregated data over municipalities as *represented variables*). Anyway, this step is easily included in our representation: this dataset is again a design matrix (Section 3), from which a further round of analyses can be considered, this time considering municipalities as statistical units.

The purpose of the presented contribution is to start defining a framework for governing metadata models through formal representations, starting from registers and ending with the dissemination step. The introduced metadata models have two specific characteristics: (i) they capture the *statistical* meaning of the artefacts produced through the production pipeline and (ii) they are thoughts as associated to layers of the statistical production process, in this way permitting to trace and maintain the *lineage* of each artefact.

In relation to how the statements in the introduction find their solution by means of the metadata strategy defined in the previous sections, these are our considerations.

- *The proposed approach allows proper statistical analyses on registers' data by providing an explicit and well-defined statistical semantics to microdata present in registers*. The distinct layers of the process allow to focus always on specific aspects: the definition of the design matrix, the definition of the desired output, the definition of the cube that includes data on the desired output. Concepts are re-used in all the layers. There are layers whose objective is to create specific queries (selection of the design matrix, selection of the features of the desired output), while others need appropriate choices (*e.g.* the *Process method*, and then the appropriate rule to consider according to the nature of the design matrix as a complete observation of the variables in the population or a sample of it).
- *The proposed approach derives output statistical datasets from statistical data cubes that, in this case as well, have an explicit and well-defined statistical semantics.* The statistical output is inserted in

data cubes where the concepts to use for *Fact*, *Measure* and *Dimension*, is once and for all clearly defined.

- *The proposed approach governs the whole production pipeline from registers' data to output datasets by means of a defined and coherent metadata asset*. This is maybe the most interesting and practical aspect to pursue. As an example, Istat has the objective to create a framework that allows users to create their own analyses from registers[3] and the work here presented is an important step in this direction.

References

Calvanese, D., G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, and R. Rosati. 2009. "Ontologies and Databases: The DL-Lite Approach". In Tessaris, S., E. Franconi, T. Eiter, C. Gutierrez, S. Handschuh, M.-C. Rousset, and R.A. Schmidt. *Reasoning Web. Semantic Technologies for Information Systems. 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30 - September 4, 2009, Tutorial Lectures*: 255-356. Berlin/Heidelberg, Germany: Springer.

Golfarelli, M., and S. Rizzi. 2009. *Data Warehouse Design. Modern Principles and Methodologies*. New York, NY, U.S.: McGraw Hill.

Lenzerini, M. 2002. "Data Integration: A Theoretical Perspective". In *PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*: 233-246. New York, NY, U.S.: Association for Computing Machinery.

Poggi, A., D. Lembo, P. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. 2008. "Linking Data to Ontologies". In Spaccapietra, S. (Ed.). *Journal on Data Semantics X. Lecture Notes in Computer Science*: 133-173, Volume 4900. Berlin/Heidelberg, Germany: Springer.

Scannapieco, M., M. Scanu, and L. Tosco. 2019. "On Designing Aggregated Data as Statistical Data Cubes". *Technical Report*. Istat Advisory Committee on Statistical Methods 2019 Meeting. Roma, Italy: Istat.

[3] This is internally managed by a dedicated working group Register-based Analytical Framework.

United Nations Economic Commission for Europe - UNECE. 2013. *Generic Statistical Information Model (GSIM): Specification*. Geneva, Switzerland: UNECE (Accessed 12th February 2023). https://statswiki.unece.org/display/gsim/GSIM+Specification.

World Wide Web Consortium - W3C. 2014. *The RDF Data Cube Vocabulary*. Cambridge, MA, U.S./Sophia-Antipolis, France/Tokyo, Japan/Beijing, China: W3C (Accessed 12th February 2023). https://www.w3.org/TR/vocab-data-cube/.

Closing SESSION

Conclusions

Linda Laura Sabbadini[1]

It is time to close this first Istat Workshop on Methodologies for Official Statistics. As you know, the Istat commitment to quality made us develop this workshop. As a producer of official statistics data, Istat needs data of good overall quality. Research on methods in official statistics is a fundamental asset for ensuring always the best methods (and hence a good quality) in the different data production contexts that Istat manages.

My personal opinion is that this event has been a success: the "market of ideas" that is typical of workshops, has been extremely live and with important feedback on the development of new methods and their implementation in the current statistical production.

For instance, the goal of the 'permanent' Census (the focus of the first session) is to produce annual data - replacing the previous decennial cycle - using information from administrative sources integrated with sample survey information. The new Census strategy is planned to allow a significant reduction of the cost of the census, of respondents' burden, and of the organisational impact on municipalities. These objectives asked for the development of new methods and tools to be used in order to sustain the whole production pipeline.

Furthermore, registers (the topic of the second session) play a central part in the Istat data production processes. Anyway, registers have often to face the problem of dealing with multiple sources of information on the variables of interest. What source should be preferred? What is the associated quality framework? What happens if one of the sources has a non-probabilistic nature?

This last aspect has been the core problem tackled in the master class provided by Professor David Haziza, of the University of Ottawa, who I would like to thank for his clarity and for suggesting so many lines along which research could be pursued.

The use of non-probabilistic samples naturally leads to the increasing role of big data and the surge of trusted smart statistics, the topic of the third

[1] Linda Laura Sabbadini, Italian National Institute of Statistics – Istat.

session. The usual statistical framework has to be completely revised. The attention of NSIs to include these sources of data among those useful for the production of official statistics can only be motivated by a clear analysis of the overall quality of the results obtainable in this context.

Finally, the last session was devoted to standardisation. This context is of extreme interest because this area is in between different topics: statistics, data science, and semantics. It touches on the areas of methods, with their documentation and implementation, as well as the area of metadata, with the efforts to make metadata coherent along a process, between the processes, and internationally for easy data exchanges and comparisons.

The wish I would like to send to all of you is that this first workshop does not stop today. On the one hand, it is important that this meeting will be renewed in the next years. Starting from these two days' experience, it would now be important to open the next workshop to talks about methods in official statistics developed elsewhere in the world by means of an open call for papers. It is time for this workshop to grow up. Secondly, I truly wish that this workshop occasion could foster possible cooperation devoted to research in the area of statistical methodology. Thirdly, I truly wish that the cooperation between research in the area of statistical methodology and statistical production in Istat becomes closer.

As a last message, I would like to thank the Advisory Committee on Statistical Methods which acted as Programme Committee of this workshop. My gratitude is also for having served in the last three years in the Committee: some of you have been working in the Committee since 2017. Your help in making research work in Istat aligned with the state of current research in statistics has been greatly appreciated.

I would like also to thank the invited discussants for the very interesting comments and remarks. Their suggestions and comments will be a motivating driver for making research more prominent in our Institute and effective for data production.